

PTVCIO-016 Data Analytics Platform

**Solution Architecture Design**

(SAD)

Approved by:

|  |  |  |
| --- | --- | --- |
| **Name** | **Signature** | **Date** |
|  |  | **\_\_\_\_\_ / \_\_\_\_\_ / \_\_\_\_\_** |
|  |  | **\_\_\_\_\_ / \_\_\_\_\_ / \_\_\_\_\_** |
|  |  | **\_\_\_\_\_ / \_\_\_\_\_ / \_\_\_\_\_** |

Document information

|  |  |
| --- | --- |
| **Item** | **Details** |
| TRIM Reference |  |
| Document Owner |  |
| Document Author |  |
| Version |  |
| Issue Date |  |

Document reviewer list

|  |  |  |
| --- | --- | --- |
| **Name** | **Position** | **Date** |
|  |  |  |
|  |  |  |

Document distribution list

|  |  |  |
| --- | --- | --- |
| **Name** | **Position** | **Date** |
|  |  |  |
|  |  |  |

Revision history

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Description** |
| V0.1 | 24-Nov-15 | Jim McLeod (CGI) | Initial draft |
| V0.2 | 17-Dec-15 | Jim McLeod (CGI) | Merged Infrastructure Design 0r31 and PTV feedback |
| V0.3 | 11-Jan-16 | Jim McLeod (CGI) | Feedback from second PTV Review |
| V0.4 | 29-Jan-16 | Jim McLeod (CGI) | Feedback from third PTV Review |
| V0.5 | 17-Feb-16 | Jim McLeod (CGI) | Feedback from fourth PTV Review |
| V0.6 | 24-Feb-16 | Jim McLeod (CGI) | Feedback from fifth PTV Review |
| V0.7 | 24-Feb-16 | Andrew Kimber (CGI) | Section 2.6 Open Items (SA-I2 & SA-I13) |
| V0.8 | 25-Feb-16 | Jim McLeod (CGI) | Additional feedback incorporated |
| V0.9.1 | 17-Jun-16 | Jim McLeod (CGI) | Updated open issues and incorporated post-EDA feedback |
| V0.9.2 | 30-Aug-16 | Jim McLeod (CGI) | Feedback from review by Karti Mahendran |
| V0.9.3 | 19-Oct-16 | Andrew C. Smith (CGI) | Section 12.4.6 – Application Whitelisting |
| V0.9.4 | 2-Nov-16 | Andrew C. Smith (CGI) | Section 12.1.1 and 12.1.2 |
| V0.9.5 | 18-Nov-16 | Andrew Kimber (CGI) | Sections 2.7, 11.5.2, 12.8, & Appendix 13.3. |
| V0.9.6 | 30-Nov-16 | Andrew Kimber (CGI) | Update WAF diagram in Section 12.1.1 |
|  |  |  |  |
|  |  |  |  |

**Table of Contents**

[1 Executive Summary 6](#_Toc468283761)

[1.1 Project Background 6](#_Toc468283762)

[**1.2** PTV Project Objectives 6](#_Toc468283763)

[1.3 Design Highlights 7](#_Toc468283764)

[1.4 Risk Assessment 7](#_Toc468283765)

[1.5 Key Architecture Decisions and Compliance 7](#_Toc468283766)

[2 Solution Context 9](#_Toc468283767)

[2.1 Related Documentation 9](#_Toc468283768)

[2.2 Key Architectural Requirements 9](#_Toc468283769)

[2.3 Assumptions 9](#_Toc468283770)

[2.4 Constraints 9](#_Toc468283771)

[2.5 Dependencies 10](#_Toc468283772)

[2.6 Open Items 10](#_Toc468283773)

[2.7 Risks 13](#_Toc468283774)

[3 Solution Overview 14](#_Toc468283775)

[3.1 Baseline (Current) Solution Overview 14](#_Toc468283776)

[3.2 Target (To-Be) Solution Overview 16](#_Toc468283777)

[3.3 Migration Plan 24](#_Toc468283778)

[3.4 Reference Architecture 25](#_Toc468283779)

[4 Business Architecture 26](#_Toc468283780)

[4.1 Business Functions/Services 26](#_Toc468283781)

[4.1.1 Changed Business Functions/Services 26](#_Toc468283782)

[4.1.2 Decommissioned Business Functions/Services 26](#_Toc468283783)

[4.1.3 New Business Functions/Services 26](#_Toc468283784)

[4.2 Business Processes 26](#_Toc468283785)

[4.2.1 Touch On Rate Surveys (TORS) 26](#_Toc468283786)

[4.2.2 Myki Data 27](#_Toc468283787)

[4.2.3 OPA Reporting 27](#_Toc468283788)

[4.2.4 MACA Reporting 27](#_Toc468283789)

[5 Data Architecture 28](#_Toc468283790)

[5.1 Conceptual Data Entities 28](#_Toc468283791)

[5.2 Logical Data Entities 31](#_Toc468283792)

[5.3 Data Flow 35](#_Toc468283793)

[5.3.1 Source to Data Lake Data Flow 35](#_Toc468283794)

[5.3.2 Data Lake Onwards Data Flow 37](#_Toc468283795)

[5.3.3 Abnormal Flow 38](#_Toc468283796)

[5.4 Data Sources 39](#_Toc468283797)

[5.4.1 Data Volumes 40](#_Toc468283798)

[5.4.2 Smartrak 40](#_Toc468283799)

[5.4.3 myki 42](#_Toc468283800)

[5.4.4 DIVA (GTFS) 43](#_Toc468283801)

[5.4.5 TransProd 43](#_Toc468283802)

[5.4.6 Transaction Rate Surveys (TORS) 43](#_Toc468283803)

[5.5 Enterprise Data Warehouse 44](#_Toc468283804)

[5.5.1 Data Vault Overview 44](#_Toc468283805)

[5.5.2 Raw Data Vault 45](#_Toc468283806)

[5.5.3 Business Vault 45](#_Toc468283807)

[5.5.4 View Layer 45](#_Toc468283808)

[5.5.5 Database Schemas 45](#_Toc468283809)

[5.6 Data Marts and Cubes 48](#_Toc468283810)

[5.7 SQL Server Integration Services 49](#_Toc468283811)

[5.8 Metadata Management 49](#_Toc468283812)

[5.8.1 Technical & Administrative Metadata 49](#_Toc468283813)

[5.8.2 Business Metadata 50](#_Toc468283814)

[5.9 Master Data Services 51](#_Toc468283815)

[5.10 Data Quality and Transformation Process 52](#_Toc468283816)

[5.11 Downstream Uses 53](#_Toc468283817)

[5.12 Future Consideration 53](#_Toc468283818)

[5.12.1 Scalability 53](#_Toc468283819)

[5.13 Data Migration Plan 53](#_Toc468283820)

[5.13.1 Metcard Data 53](#_Toc468283821)

[5.13.2 Bus Tracking System Data 55](#_Toc468283822)

[5.13.3 myki Ticketing Data 55](#_Toc468283823)

[6 Information Architecture 56](#_Toc468283824)

[6.1 Reports and Formats 56](#_Toc468283825)

[6.1.1 SQL Server Reporting Services 57](#_Toc468283826)

[6.1.2 Power View 57](#_Toc468283827)

[6.1.3 Excel Services 57](#_Toc468283828)

[6.1.4 Power BI 57](#_Toc468283829)

[6.1.5 ADA Based Reporting 57](#_Toc468283830)

[6.2 Dimensional Matrix 58](#_Toc468283831)

[6.3 Data Dictionary and Business Rules 59](#_Toc468283832)

[7 Application Architecture 60](#_Toc468283833)

[7.1 Data Lake - HDInsight 60](#_Toc468283834)

[7.1.1 Data Lake Storage Accounts and Containers 61](#_Toc468283835)

[7.2 SQL Server Analysis Services – Semantic Layer 61](#_Toc468283836)

[7.3 SQL Server Analysis Services Server Infrastructure 62](#_Toc468283837)

[7.3.1 Disk Storage Requirements 62](#_Toc468283838)

[7.3.2 Memory and Processor 62](#_Toc468283839)

[7.4 Analytics Discovery Area 63](#_Toc468283840)

[7.4.1 Virtual Machines 63](#_Toc468283841)

[7.4.2 HDInsight (Hadoop) 64](#_Toc468283842)

[7.4.3 Azure SQL Database (Sandpit) 64](#_Toc468283843)

[7.4.4 SQL Server Analysis Services 64](#_Toc468283844)

[7.4.5 Spark 64](#_Toc468283845)

[7.4.6 Azure Machine Learning 65](#_Toc468283846)

[8 Integration Architecture 66](#_Toc468283847)

[9 SharePoint 2013 Architecture 67](#_Toc468283848)

[10 Presentation (Reporting) Architecture 69](#_Toc468283849)

[10.1.1 Reporting Presented through SharePoint 2013 69](#_Toc468283850)

[10.1.2 Power BI 70](#_Toc468283851)

[11 Technology Architecture 73](#_Toc468283852)

[**11.1** Environments 73](#_Toc468283853)

[11.2 Major Software Versions and Licensing 74](#_Toc468283854)

[11.2.1 Licensing for Non-Production 76](#_Toc468283855)

[**11.3** Network View 76](#_Toc468283856)

[11.3.1 Connectivity 76](#_Toc468283857)

[11.3.2 Data Transfers 78](#_Toc468283858)

[11.3.3 Network Topology 79](#_Toc468283859)

[11.3.4 Network Addressing 80](#_Toc468283860)

[11.3.5 Network Services 81](#_Toc468283861)

[11.3.6 Network Security Groups 81](#_Toc468283862)

[**11.4** Infrastructure Architecture 81](#_Toc468283863)

[11.4.1 Overview 81](#_Toc468283864)

[11.4.2 Resource Groups 82](#_Toc468283865)

[11.4.3 Azure PaaS 82](#_Toc468283866)

[11.4.4 Azure IaaS 83](#_Toc468283867)

[11.4.5 Azure Storage 83](#_Toc468283868)

[11.4.6 Azure Storage Accounts 83](#_Toc468283869)

[11.4.7 Availability Sets 84](#_Toc468283870)

[11.4.8 Active Directory – DAP 85](#_Toc468283871)

[11.4.9 Active Directory - PTV 86](#_Toc468283872)

[11.4.10 Azure Resources 87](#_Toc468283873)

[11.4.11 Future Infrastructure Scalability Options 92](#_Toc468283874)

[**11.5** Operations Architecture 93](#_Toc468283875)

[11.5.1 Monitoring and Event Management 93](#_Toc468283876)

[11.5.2 Simple Mail Transfer Protocol (SMTP Email) 93](#_Toc468283877)

[11.5.3 Performance Management and Capacity Management 94](#_Toc468283878)

[11.5.4 Data Protection 94](#_Toc468283879)

[11.5.5 Disaster Recovery 96](#_Toc468283880)

[11.5.6 Disaster Recovery for Data Services 97](#_Toc468283881)

[11.5.7 Business Continuity 97](#_Toc468283882)

[11.5.8 Patching, Maintenance, Software Releases 97](#_Toc468283883)

[12 Security Architecture 99](#_Toc468283884)

[12.1 Network Security 99](#_Toc468283885)

[12.1.1 Web Application Firewall (WAF) 100](#_Toc468283886)

[12.1.2 Access Control Intermediary (Proxy) for ADA environment 101](#_Toc468283889)

[12.2 System Security 102](#_Toc468283890)

[12.2.1 Azure Subscription Access 102](#_Toc468283891)

[12.2.2 Virtual Machines Management 102](#_Toc468283892)

[12.2.3 Logging 103](#_Toc468283893)

[12.3 Data Security 103](#_Toc468283894)

[12.3.1 Data Lake - HDInsight 103](#_Toc468283895)

[12.3.2 Azure SQL Database and Azure SQL Data Warehouse 104](#_Toc468283896)

[12.3.3 Power BI Data 105](#_Toc468283897)

[12.3.4 Azure Machine Learning 105](#_Toc468283898)

[12.3.5 SQL Server Analysis Services – Semantic Layer 105](#_Toc468283899)

[12.3.6 Storage Accounts and Containers 105](#_Toc468283900)

[12.4 Application Security 106](#_Toc468283901)

[12.4.1 SharePoint Authentication 106](#_Toc468283902)

[12.4.2 SQL Server Analysis Services (SSAS) 107](#_Toc468283903)

[12.4.3 SQL Server Integration Services (SSIS) 107](#_Toc468283904)

[12.4.4 Master Data Services (MDS) 107](#_Toc468283905)

[12.4.5 Certificates 107](#_Toc468283906)

[12.4.6 Application Whitelisting 107](#_Toc468283907)

[12.5 User Client Security 108](#_Toc468283909)

[12.5.1 DAP AD Account Management 108](#_Toc468283910)

[12.5.2 Source Code Access 109](#_Toc468283911)

[12.6 User Classes 109](#_Toc468283912)

[12.7 User Permissions Matrix 110](#_Toc468283913)

[12.8 User Maintenance Processes 111](#_Toc468283914)

[12.8.1 Internal PTV Users 111](#_Toc468283915)

[12.8.2 External Users 111](#_Toc468283916)

[12.8.3 User Onboarding/Offboarding Overall Process Flow 113](#_Toc468283917)

[12.8.4 User Onboarding/Offboarding Sub-Process Flow 114](#_Toc468283918)

[13 Glossary, Appendix & Attachments 115](#_Toc468283919)

[13.1 Glossary 115](#_Toc468283920)

[13.2 Appendix 115](#_Toc468283921)

[13.3 Attachments 115](#_Toc468283922)

# Executive Summary

## Project Background

The PTV Corporate Plan 2014-2018 identifies six strategic themes that will guide PTV's business. One of these is Information, Systems and Processes. A key commitment in this theme is to establish a single source of information for PTV employees which can be easily accessed from commercial off-the-shelf systems. The DAP is an initiative central to the realisation of this commitment.

The project will commence with a Proof of Value (POV) phase that will be limited to achievable outputs that will quickly demonstrate the value of the DAP to key stakeholders including the PTV Executive and the PTV Board, therefore, the POV is focused on automating and delivering reporting practices for patronage and revenue, bus operational performance and an analytics sandpit capability.

In the future, the DAP is likely to include the reporting of customer experience and integration/replacement of the Transport Information Online service (operational performance of all public transport modes).

## PTV Project Objectives

PTV currently processes data from a number of sources, and this includes data about public transport timetables and network topology, vehicle telemetry data, ticketing data, and passenger counts. This data is processed within PTV by a variety of legacy systems to support the analysis and reporting activity of several PTV business units. Within PTV, the business processes for the ingestion, storage, and processing of data and the reporting of information, both internally and externally, have developed over many years as a number of separate transport organisations evolved and were merged to form PTV.

It is envisaged that the DAP will be an enduring capability/service within PTV for many years, and the specific services it supports will extend and evolve over time.

Whilst the support of PTV reporting practices is an immediate priority, the ability of the DAP to support prediction of transport network performance, patronage and revenue is a longer term goal.

It is expected that the Data Analytics Platform implementation will deliver an enterprise-wide analytics platform over a series of phases to achieve the following outcomes:

Phase 1 will be a successful Proof of Value that will:

* Initially addresses the labour intensive production of patronage and revenue reporting.
* demonstrate it can replicate and build on the success of Transport Information Online for Bus Operational Performance,
* provide an analytics sandpit,
* and that this capability can be taken to the broader business.

The main business benefits to be derived from the DAP will include:

* provide authoritative sources of information which will support PTV's reporting functions and inform business decision making processes
* reduce the manual effort needed to produce reports
* increase the range of business metrics that PTV can report on
* improve visibility into the reporting of PTV's business performance
* improve the accuracy of reporting with reduced risk of error
* increase the frequency and timeliness of reporting activity
* deliver enhanced visualisation of reporting and statistical dashboard capabilities
* support a significant (exponential) growth in the quantity of disaggregated reporting
* provide a robust and extensible capability for PTV's future predictive analytics needs

Given the anticipated longevity of the DAP, PTV are seeking a solution to not only support current PTV business practices and reporting activity, but is flexible and extensible enough to adapt to and support PTV's future business needs.

## Design Highlights

The main highlights of the design described in this document are:

* PTV will have a modern BI solution, built on the latest version of the Microsoft BI stack
* A unified view of PTV’s data assets will be compiled into a data warehouse which utilises the Data Vault methodology, designed to adapt to structured and unstructured data
* An extensive cloud-based environment with low lead times for adding capacity, and consumption charges at competitive levels

## Risk Assessment

An initial risk assessment has been performed, and is available on the PTV DAP project site hosted on the CGI Partner Workspace.

As the DAP does not initially remove any functionality that PTV staff depend on for their day-to-day tasks, implementation of the DAP has a low impact.

## Key Architecture Decisions and Compliance

The following key architecture decisions have been made:

* The DAP solution is built on the standard Microsoft BI platform, consisting of a SharePoint 2013 (IaaS) presentation layer. The web-based components are supported by Microsoft on all major browsers. SharePoint Online (Office 365) is not suitable for the DAP due to its lack of advanced reporting and deep analytics functionality, and restrictions in features external users can perform (such as self-service BI using Power View).
* Microsoft Power BI has been chosen as the preferred self-service BI tool. Power BI is supported on Windows Mobile, Android and iOS.
* The DAP environment is hosted in the Microsoft Azure cloud, located in the Australia Southeast (Melbourne) data centre, with data geo-replicated to Azure Australia East (Sydney).
* The DAP is available to external (non-PTV) users, with the exception of Power BI reports, as Power BI is only available within a single organisation and reports cannot be shared with external parties. To offset this, SharePoint offers self-service BI for external users through Power View and Excel Services.
* The solution is scalable with respect to concurrent users and increasing increased data volumes. As there are many different components in the DAP, each may scale in different ways. See section 3.2 Target (To-Be) Solution Overview for more details on each component. At a high level, Azure Platform as a Service and Software as a Service options have been chosen where possible, falling back to Infrastructure as a Service for those components where a cloud option is not yet available. While the DAP is scalable to support larger loads, this is not an automatic activity.
* In order to meet the requirement for communication should be performed on secure channels, the solution will tunnel through SSL encrypted connections: HTTPS for web traffic, and FTPS for file transfers (File Transfer Protocol Secure – not Secure Shell FTP). FTPS was chosen over S-FTP due to the native support on Windows Servers for FTPS.
* The Proof of Value DAP has been designed to not require direct database access from the GSP network as the SharePoint front-end provides all static reporting functionality, Power BI provides self-service BI, and the ADA environment is hosted within the cloud and accessed via HTTPS. This decision was taken based on advice from PTV that direct access to SQL sources in Azure would not be permitted from within the PTV network.
* The data model provided within the EDW is based upon industry standard terminology (TransModel) to describe public transport entities. These entities will form the basis of, or map back to PTV’s Enterprise Information Model.
* All sourced data is to be stored within a Data Lake that can allow ad hoc analysis, with targeted sub-sets of data being cleansed and loaded into an Enterprise Data Warehouse for reporting and data exploration.
* The Enterprise Data Warehouse (EDW) is modelled using the Data Vault methodology to allow easy integration of new data sources in future. The options for EDW were: a) Data Vault, b) classic Inmon (3rd normal form Data Warehouse), and c) Kimball (collection of dimensionally-modelled data marts). Data Vault was chosen over the other options due to the flexibility in adding new data sources and functionality, and to provide one standard view of PTV’s data – an important and necessary outcome as the majority of the DAP Proof of Value data sources are not accessible via a direct database connection from the Data Warehouse. Bill Inmon stated in 2007 that the Data Vault is the preferred modelling technique for his concept of the Data Warehouse in modern data warehouse systems.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Inmon** | **Kimball** | **Data Vault** |
| **Staging** | Y | Y | Y |
| **3rd Normal Form EDW** | Y | - | Y |
| **Data Marts** | Y | Y (persisted) | Y |
| **Cubes** | Y | Y | Y |

* The Data Marts are modelled using a Kimball methodology to provide optimised reporting.
* The solution provides an Analytics Discovery Area (ADA) to provide users with an analytics sandbox where data discoveries and additional data enrichment can be performed. This area provides a number of tools that can be used by PTV’s data scientists. As agreed with project SMEs, the ADA environment is considered a “production sandpit” environment, with no need for ADA Development and Test environments. No ADA functionality is required for the core data warehouse as described in this document. Some components produced by ADA users may require a separate environment to “Productionise” those components, but infrastructure for these can be determined as needed; this is a potential future requirement and is not within scope for the proof of value.
* The DAP solution has been designed such that components do not initiate inbound connections back into the PTV/CenITex environment. All data sources are thus designed as “push to cloud” using a file-based mechanism into the Data Lake. The DAP FTPS service resides on a bastion host to provide an additional layer of security.
* For the future, Azure Data Lake (ADL) is an upcoming solution for the DAP for PTV’s Data Lake as it integrates Active Directory security and improves big data analytics, but it is still in preview, and not expected to be available in Australia until Q3 2016 at the earliest (June 2016). The Proof of Value DAP will still meet requirements using HDInsight and Windows Azure Blob Storage to provide a Data Lake, and the upgrade path from a HDInsight/Windows Azure Storage Blob (WASB) Data Lake to ADL is clear as ADL builds on top of HDInsight and WASB, with tools to migrate from WASB to ADL.
* Data Lake options considered for DAP were: a) HD Insight, b) Azure Data Lake, c) Windows Azure Storage Blob (WASB) Data Lake. As the requirement for Data Lake is storage and the SQL Data Warehouse accesses the DL storage directly, the decision was made to implement option (c) Windows Azure Storage Blob, which can be presented to HD Insight clusters, including to ADA. Option b) Azure Data Lake, whilst desirable as a future roadmap item, is not available in the required timeframe for PoV and was not included in CGI’s Offer for PoV.

# Solution Context

## Related Documentation

The following documents were used as sources of information for this Solution Architecture Design:

* “PTV Services Interface Design Document: Processed Transport Data Exchange (PTDE)”
* PTV DAP RFT Non-Functional Requirements - attached to this document in section 13.3.

The following future documents will supplement this document:

* PTV DAP Developer Guidelines (includes technical naming conventions)
* PTV DAP SharePoint User Interface Design
* PTV DAP Data Mapping Specifications
* PTV DAP Reporting Master Specification
* PTV DAP Reporting Specifications (per set of reports)
* DAP Security Assessment Report

## Key Architectural Requirements

The following key architectural requirements for the solution have been taken into account when developing the solution design:

* All solution data must be stored in Australia.
* The solution must maintain discrete production and non-production systems.
* System must support no less than 99.5% availability. The Respondent must provide details of the redundancy set-up, auto-failover within the primary environment, and backup architecture within the System to support its availability compliance.
* The solution must backup all PTV Data, at a minimum, on daily basis.
* Ability to keep the solution operational by swapping to a failover site within one hour of the main site failing
* The solution must have sufficient fault tolerance that in the event of a catastrophic system failure, the data lost will be no more than 20 minutes’ worth of operation.
* The solution must monitor the health of the systems and detect incidents to enable prompt resolution for all services.
* The overall solution should be able to be accessed in its entirety through standard web browsing ports (i.e. Port 80 and Port 443) so that users both within the PTV corporate network, and outside the network on a standard desktop can obtain secure client access.

## Assumptions

| # | Assumption | Description |
| --- | --- | --- |
| SA-A1 | Existing PTV network infrastructure meets requirements | It has been assumed that PTV’s existing network infrastructure has sufficient capacity to handle any extra load generated by PTV staff accessing the DAP. |

**1: Assumptions**

## Constraints

| # | Constraint | Description |
| --- | --- | --- |
| SA-C1 | Outbound connectivity limited to HTTP and HTTPS | All outbound connectivity from PTV’s internal network is limited to HTTP and HTTPS ports. |

: Constraints

## Dependencies

| # | Dependency | Description | Status | Who | When |
| --- | --- | --- | --- | --- | --- |
| SA-D1 | Provision of services by CenITex | The solution as designed relies on several services that are either supported by or provisioned by CenITex. These include:   * Outbound internet connectivity (Internet and ExpressRoute); * The synchronisation of PTV Active Directory accounts to Azure AD; * The authentication of PTV Azure AD accounts using federation services.   These services are already implemented (Internet connectivity and ADConnect for synchronisation of PTV AD accounts to Azure AD), or available as a CenITex service offering (Tivoli FIM authentication), or coming soon (ExpressRoute). | Cenitex raising MSR, work not started yet. | JK/KD | ExpressRoute 1/6/16  Authentication29/02/16 |
| SA-D2 | Fujitsu to develop TransProd extract and send via FTPS | The TransProd database is hosted and maintained in Fujitsu’s data centre, and the consensus is that Fujitsu should be engaged to write the extraction code to CSV based on our specifications. | In progress | KD/AK | March 2016 |
| SA-D3 | NTP and HOBAN will redevelop their TORS results in CSV and upload to FTPS server | If NTP and HOBAN are redeveloping their TORS spreadsheets to PTV’s new CSV format specification. | In progress | KG | May 2016 |

**3: Dependencies**

## Open Items

| # | Issue | Description | Status | Who | | When | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SA-I1 | DAP authentication against CenITex Tivoli FIM | CenITex currently have an MSR in progress to provide an architecture for the DAP’s authentication mechanisms back to the GSP network. CenITex have implemented single sign-on to Azure services internally, and are sure the solution is available for PTV. Closure: SAML Authentication from SharePoint to Tivoli FIM has been defined and implemented. | Closed | JK/KD | | 29/02 | |
| SA-I2 | Auto scaling | PTV Application Architect has requested a solution for auto-scaling of the DAP SharePoint environment. This is a change to the project scope, as it is not a stated required and was not included in CGI’s Offer. Closure: As agreed with PTV DAP Project Manager, this will be investigated as a requirement for a future phase. | Closed for PoV Phase | EDA | | 25/02 | |
| SA-I3 | Express Route | CenITex are in the process of implementing an Express Route connection. Deemed a pilot project, there may be some issues getting Express Route running, and it is not expected to be active until April-May 2016. Until then, the DAP can use the Internet connection, but performance may be impacted by other government uses of the WoG Internet connection. Closure: This is closed from an architectural point of view, pending completion of the ExpressRoute provisioning. | Closed | JK/KD | | 01/06 | |
| SA-I4 | Power BI Location | Power BI is currently hosted in the US, and can support both DirectQuery (data is queried from the US Power BI server, report rendered and sent to the user), or uploading of static data sets (refreshed once every 24 hours). Closure: Power BI is now available hosted in Australian Azure Data Centres. | Closed | CGI | | 02/16 | |
| SA-I5 | Azure Machine Learning Processing Location | Azure ML is currently not available in the Australian Azure regions, meaning data must be transferred for processing to the US Azure servers. Subject to security assessment, this can be mitigated by only choosing to access data sets that are unclassified and approved for processing outside of Australia (removing any sensitive data), or by performing machine learning within the ADA environment, such as on the Spark cluster. Closure: The business is aware of this limitation, and is prepared to wait until AML is available locally. | Closed for PoV Phase | | CGI | | 16/03 | |
| SA-I6 | Outstanding Data Classification | Some data has not been formally classified. The absence of classification is not deemed to be a problem, but classification should occur sooner rather than later.  **Closure:** This is an on-going task separate from the DAP. PTV has agreed that data to be stored for PoV is acceptable to be held in Australian Azure servers, with the encryption of key columns as outlined in section 12.3.1. | Closed | | PTV | |  | |
| SA-I7 | Dimensional Matrix not yet determined | The Data Mart designs are not done until the Reporting Specification. Data sources are correctly identified as supplying the data required, but Reporting Specifications will be done in the next 4-8 weeks, solidifying the Dimensional Matrix. Closure: This is now complete, and the Dimensional Matrix updated in section 6.2. | Closed | | CGI | | 21/03 | |
| SA-I8 | SMTP Server required for alerting | Exact nature of the SMTP server has not yet been determined – whether a CenITex SMTP server is an option, an Azure SaaS option, or self-hosted. Closure: SendGrid Azure email service is to be used – details in section 11.5.2 | Closed | | CGI | |  | |
| SA-I9 | RTO Clarification | RTO listed as 1 hour, but not specified whether this is 1 hour from decision to failover, 1 hour from start of incident, or 1 hour from time incident is opened in ITMS. Closure: RTO will commence from the point where the incident is first reported. | Closed | | KD | | 01/03 | |
| SA-I10 | Myki extract has a manual component | As there is no direct access to the myki data warehouse, the solution has been designed to operate using an automated extract that is manually started and uploaded. This will ideally change when NTT are able to commit time to assisting with making the data more accessible. See Chris Maloney. | Closed for PoV Phase | | Chris Maloney | |  | |
| SA-I11 | Is GTFS the best source of timetable data as opposed to querying DIVA directly? | Question raised by Vikas. GTFS is used as it is the public-facing timetable information, and is reviewed prior to release. Closure: The project has determined that GTFS mixed with TransProd data is the best source for the Proof of Value phase. This may change in the near future as Timetables On Demand becomes in use. | Closed | | RB/DA | | 25/02 | |
| SA-I12 | Security Architecture Review | PTV is arranging a review of the security architecture of the DAP, which will include a review of Section 12, and the decision around a single Virtual Network with NGS between Dev/Test/Prod layers versus multiple Virtual Networks. Closure: Security architecture review has been completed, and Section 12 updated in this document. | Closed | | PTV | | 1/04 | |
| SA-I13 | ADA environment access | The DAP ADA environment is located in Azure and requires access via remote desktop services. This will require the ADA user to enter a username and password to initiate their ADA session. The project has previously agreed to this approach. Microsoft has a potential solution to streamline this access, however this solution is in preview. Closure: As agreed with the project, this direction is acceptable for the current time and can be addressed in a future phase. | Closed for PoV Phase | | JK | | 26/2 | |

**4: Issues**

## Risks

| # | Risk | Likelihood | Owner | Severity | Potential Impact/Mitigation Strategy |
| --- | --- | --- | --- | --- | --- |
| SA-R1 | The PTV DAP domain in the cloud (dap.int) may not be able to authenticate PTV users for SQL Server Analysis Services. | Low | KC | High | Authenticate users against CenITex Tivoli service, or synchronise users into the dap.int domain. |
| SA-R2 | Complexity of PTV infrastructure may impact delivery of DAP (Azure/PTV/CenITex/Government) | Low | KC | High | Ensure internal infrastructure is well understood before producing design. Meet with PTV infrastructure management staff during build to discuss any potential issues. |
| SA-R3 | IT network performance may impact response time | Low | KC | High | Provide estimated usage statistics to PTV, monitor usage during initial deployment stages. |
| SA-R4 | General Availability of Azure SQL Data Warehouse date slips from Q1 2016 | Low | JMc | Medium | Decision to be made whether EDW goes live using SQL DW in preview mode, or redeployed on IaaS SQL Server 2014. |
| SA-R5 | Address environment segregation (NEW) | Low | AK | High | Potential for Dev/Test environments to tamper with Production data.  Risk assessment needed.  Refer to attached Risk Assessment – see Appendices 13.3 below. |

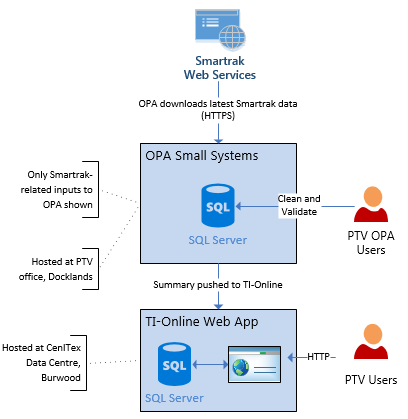
: Risks

# Solution Overview

## Baseline (Current) Solution Overview

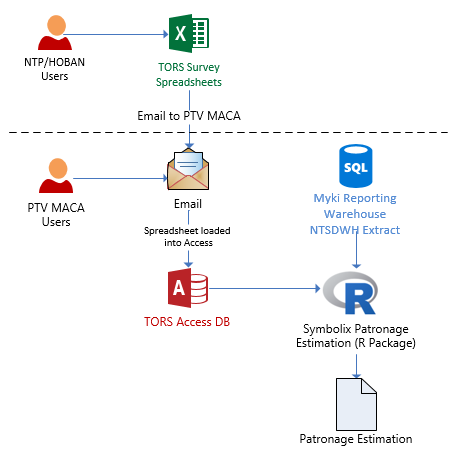
There is no one solution currently in place that the proof-of-value DAP will replace – it is multiple systems and processes.

The OPA team currently load Smartrak bus data into a SQL Server 2008 database, process and clean the data, and upload to the TI-Online web application.



Current Solution – Smartrak Bus Data

MACA utilise spreadsheets, Access databases and direct database queries to analyse data.



Current Solution – Patronage Estimation Calculation

## Target (To-Be) Solution Overview

The DAP solution is a Microsoft Business Intelligence solution hosted completely within the Microsoft Azure cloud, located in the Australian Azure Data Centres.

The primary aim of the platform is to load in data from multiple data sources, from multiple areas of the business and combine the data sources into one unified view of the organisation as much as possible. It won’t always be possible to have one single view, as the same facts can show different sort of truths, depending on the business unit that is viewing the data. For example, service performance may be acceptable from a contractual point of view, but fail from a customer experience and perception point of view.

An overview of the proposed solution is shown below.

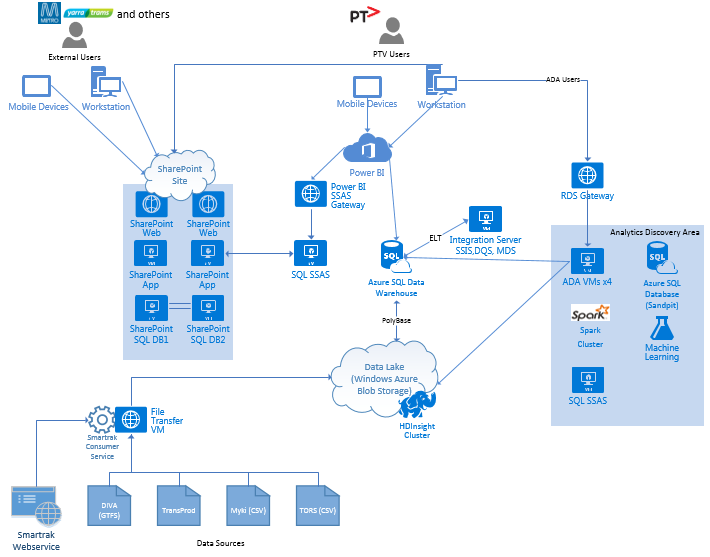


Solution Overview

The major components shown in this diagram are:

* **Data Sources** – None of the data sources are available directly to the DAP, and must be first brought into the DAP environment before processing.
* **Data Lake** – The Data Lake is an organised repository of data sources with the philosophy of all data may be useful in the future. The Data Lake is automatically extensible to petabyte size.
* **SQL Data Warehouse** – The Enterprise Data Warehouse (EDW) is an organised repository of cleansed data that has been validated and enriched and is the main source of facts about PTV’s business.
* **Analytics Discovery Area** – The ADA provides a sandpit environment where data scientists can query data stored in the DAP, create models, and seek insights from the entire set of data using big data and statistical analysis tools. The tools provided are:
  + **HDInsight** – Microsoft’s Hadoop implementation for big data queries from disk
  + **Spark (on HDInsight)** – Microsoft’s Spark implementation for big data analytics and machine learning based on the Spark platform.
  + **SQL Databases** – Providing a database environment for cleansing and processing data as part of analytics.
  + **Custom Analytics Software** –R Studio, Python and SQL Server tools, these tools enable the data scientist users of the ADA environment to create the analyses they want, using the tools that most suit the purpose.
  + **Azure Machine Learning** – A 100% cloud-based Machine Learning suite from Microsoft that provides an easy interface to creating Machine Learning experiments.
* **SQL Server Analysis Services Cubes** - The Proof of Value DAP provides three analytics cubes:
  + Modal Revenue Estimates
  + Modal Patronage Estimates
  + Bus Service Performance
* **SharePoint** - The SharePoint portal provides internal and external users to a **curated** portal to view reports. There are different sections available, including Patronage, Revenue and Operational Performance that users will be granted rights to view, depending on their requirements and privileges.
* **Power BI** – Power BI provides a self-service BI capability allowing internal PTV users to explore datasets presented by the EDW and the Analysis Services cubes, create reports, and share them internally. Power BI is a very new technology, and deep reporting features are not currently available – these kind of reports will be presented through SharePoint.
* **Users** - There are three main classes of user, with each class having varying permissions inside each of the areas they can access:
  + PTV/Internal Users – internal users, including users from other government departments. These users will have access to the SharePoint portal, and to Power BI (PTV internal users only). Access to the various SharePoint sites is controlled based on the user’s role in PTV.
  + External Partners – such as Metro Trains, Yarra Trams, Researchers. These users will have access to the SharePoint portal.
  + ADA Users – users of the ADA environment. These users will primarily be internal PTV users, but may include external research partners. These users will also be Internal or External SharePoint users, but as part of their ADA role they have access to various layers of data depending on their access level, , including the EDW, analytics cubes, and the Data Lake, as well as the ability to run analytics tools (such as R Studio, Spark) within the Azure environment without needing to load the data back into the PTV network.

An overview of the proposed solution is shown below.



Technical Solution Overview

Note that the diagram above shows the Production design, and separate Development and Test environments also exist. Some infrastructure is shared between regions, namely the domain controllers and remote desktop gateways.

The following table describes each major component, and each is described in the rest of this document.

|  |  |  |
| --- | --- | --- |
| Component | Category | Description |
| SharePoint 2013  (IaaS) | Reporting | An IaaS deployment of SharePoint 2013 used to provide the primary user interface. SharePoint consists of:   * A load balanced end point * Two Front End Web servers * Two Application servers, running SQL Server Reporting Services * Two SQL Server 2014 Database servers configured as a Availability Group on a Windows Server Failover Cluster   SharePoint presents static reports through SQL Server Reporting Services, as well as reports on the DAP data processing health, and serves as an interface to master data.  **IaaS**  SharePoint 2013 is an IaaS deployment of the on-premises version of SharePoint, as SharePoint Online (Office 365) does not support SQL Server Reporting Services and other analytical features that are required to meet the DAP requirements. SharePoint Online does not allow external users to use Power View, PowerPivot or Excel Services. See <https://technet.microsoft.com/en-us/library/dn635309.aspx> for information on why SharePoint on Azure Infrastructure Services is chosen.  **Scaling Options**  As mentioned in the link above, dynamic machine allocation (auto scale) is not supported with SharePoint 2013 in Azure. The farm and the VM specifications have been designed as a medium-sized deployment, capable of supporting 1000 users (Medium farms can support 10,000 users, but reporting and analytics increases the load compared to document storage).  For more information on SharePoint sizing, see <https://technet.microsoft.com/en-us/library/cc262485.aspx>.  If performance testing shows that the designed farm is not sufficient to meet the load, individual servers will be scaled up to provide more CPU cores and RAM (no licensing implications). If scaling up cannot provide the performance required to meet target response times (2 seconds response times for normal operations, 20 seconds for report generation), additional servers will be built and added to the farm, to come online at specific times of stress in response to high CPU or at scheduled times (with additional licensing implications).  **Redundancy**  Redundancy is provided by two copies of each server (Web, Application, Database), each in Availability Sets. As virtual machines in Availability Sets, Azure will ensure at least one machine is available, and will automatically restart machines on a new host if needed. |
| Azure SQL Data Warehouse  (DBaaS) | Data Processing | Azure SQL Data Warehouse is a massively parallel data warehouse cluster running on the Azure SQL platform.  **Scaling Options**  Azure SQL Data Warehouse is scaled by changing the number of Data Warehouse Units available to the Data Warehouse. The Data Warehouse’s storage is divided into 60 shards (this value is set by Microsoft and cannot be changed), and when started or scaled, Azure attaches a proportion of these shards to a number of servers. For example, 200 DWU provides two worker nodes, each with 30 shards, while 1000 DWU provides 10 nodes, each with 6 shards. At the current time, there is no auto-scale option, but the sizing of the data warehouse will be scripted and scheduled. Each multiple of 100 DWU incurs a cost of approximately $1.70 per hour.  This scaling must be scheduled, as there is a delay of 60 seconds during which the data warehouse is paused while resizing. This operation is not desirable during times when users are on the system.  To prevent any one user from bringing the Data Warehouse’s performance down, each user is assigned to a resource class, indicating how much of the Warehouse’s resources that one user can use. This allows throttling of ad hoc users, and will prevent ad hoc queries from interfering with Portal reporting users.  **Redundancy**  Redundancy is provided as part of Microsoft’s DBaaS. If a server goes down, Azure will automatically bring in another node from the Azure pool of Azure SQL Database hosts. |
| PolyBase | Data Processing | PolyBase is an interoperability layer allowing Azure SQL Data Warehouse to communicate with different data sources, presented to the Data Warehouse as External Tables. In the DAP’s case, PolyBase allows the Data Warehouse to open compressed files in the Data Lake for processing. |
| SQL Server Analysis Services (SSAS)  (IaaS) | Data Processing | SSAS provides multidimensional and tabular analytics to provide summarised access to data presented from the Data Warehouse. SSAS is not cloud aware, and is considered an “on-premises deployment in the cloud”.  **IaaS**  There is currently no cloud version of SSAS, hence the decision to use an IaaS implementation.  **Scaling Options**  SSAS can have the individual VMs scaled up, or can have multiple identical (read only) servers hosted in a farm behind a load balancer. No automatic scaling option is provided for DAP PoV – the decision to scale up or add additional servers (with additional consumption-based compute costs) will be made when necessary.  **Redundancy**  All cubes hosted in SSAS are duplicated onto the ADA SSAS environment, with a load balancer directing traffic to the Production SSAS environment. If an issue occurs where the Production SSAS instances are not available, the load balancer will redirect traffic to the ADA SSAS instances. |
| SQL Server Reporting Services (SSRS)  (IaaS) |  | SSRS is used for **static** reports, hosted and presented through SharePoint. Reports in SharePoint are accessible to desktop and mobile devices through a web browser interface.  **IaaS**  There is no equivalent cloud-service for SSRS that can provide the same level of detailed reporting that SSRS can provide to meet the DAP requirements.  **Scaling Options and Redundancy**  As SSRS is hosted on the SharePoint servers, the same rules apply as SharePoint. |
| Power BI  (PaaS) | Reporting | Power BI is a cloud-based reporting, analytics and dash boarding tool. Power BI is used for **ad hoc, self-service BI reporting**. Reports in Power BI are accessible to desktop and mobile devices through a web browser interface, or through an application (Power BI Desktop for Windows, native applications for iOS, Android, Windows 10 Phones). |
| Integration Server  (Iaas) | Data Processing | The Integration Server runs SQL Server Integration Services, Data Quality Services and Master Data Services, and is the primary server responsible for co-ordinating and scheduling data loads and scheduled tasks.  **IaaS**  Microsoft provide two options for loading data – SSIS (IaaS) and Azure Data Factory (ADF, SaaS). ADF is designed for big data loads and is automatically scalable. However, as it is new (initially released in 2015), it does not yet have the number of advanced transformations required by the DAP that SSIS provides, and the bulk of the data loaded into the DAP is not at the scale where ADF becomes necessary. Additionally, developer tool support is not as advanced as SSIS (ADF frequently requires writing JSON-based commands), hence the decision to use SSIS as an IaaS implementation.  **Scaling Options**  SSIS servers will be scaled up to provide better performance, if required. If this is not sufficient, a secondary SSIS server will be created and the workload shared across the two servers (manual configuration – SSIS does not have the concept of a cluster).  **Redundancy**  SSIS is hosted on a standalone server. In the event of a server failure that cannot be recovered by Azure automatically migrating to a new VM host, the server will be restored from backup. During this time, new data loads (currently overnight only) will not occur, but the rest of the DAP will be available. |
| Power BI SSAS Gateway  (IaaS) | Data Processing | The Power BI SSAS Gateway provides an endpoint for Power BI to access on-premises SQL Server Analysis Services cubes and tabular models (as mentioned above, SSAS is considered an “on premises” application.  **IaaS**  The Power BI SSAS Gateway is a gateway that allows Power BI to connect to an IaaS SSAS instance, and is thus firmly in the IaaS space. |
| Data Lake  (PaaS) | Data Processing | The Data Lake is a repository for all structured and unstructured data not being stored in a DBMS. Data flows into the Data Lake from each data source, where it is either processed into the Data Warehouse, left available for analytics at a later date, or both.  Cleansed and structured data can be extracted from the Data Warehouse and left in the Data Lake to provide a warm copy of historical data, accessible via PolyBase.  **Scaling Options**  See the Windows Azure Storage row in this table for Data Lake scaling options. |
| HDInsight (Hadoop)  (PaaS) | Analytics | The HDInsight cluster provides a Hadoop-based cluster to support the Data Lake. HDInsight is Microsoft’s PaaS implementation of Hortonworks Data Platform and is 100% compatible with HDP.  **PaaS**  Hadoop is available in Azure as either a PaaS implementation (HDInsight) or an IaaS implementation (manual installation and maintenance of a Hadoop cluster). The PaaS implementation (HDInsight) was chosen to reduce the maintenance effort, and to put the availability of the cluster nodes back on Microsoft. Additionally, accessing the Data Lake’s Windows Azure Blob Storage on the cluster is easier as HDInsight provides an HDFS interface from the cluster nodes to the Azure Blob Storage.  **Scaling and Redundancy Options**  As a PaaS platform, Microsoft controls the availability of the HDInsight cluster. At time of writing, the cluster is not scalable, but configuration is stored inside an Azure SQL Database, and can be destroyed and recreated at a larger size easily. |
| File Transfer VM  (IaaS) | Data Processing | The File Transfer VM is a bastion host providing a gateway for files to be transferred to the DAP for processing. It runs an FTPS server, and transfers uploaded files into the Data Lake at scheduled intervals.  **Scaling Options**  While the File Transfer VM has been sized for the requirements and should not need further scaling, it will be scaled up with little effort if required.  **Redundancy**  The File Transfer VM is hosted on a standalone server. In the event of a server failure that cannot be recovered by Azure automatically migrating to a new VM host, the server will be restored from backup. During this time, new data uploads will not be possible, but the rest of the DAP will be available. |
| Smartrak Consumer Service  (Custom Software) | Data Processing | The Smartrak Consumer Service is a Windows service running on the File Transfer VM that queries the web services provided by Smartrak to download bus tracking data into the DAP. |
| RDS Gateway  (IaaS) | Infrastructure | The Remote Desktop Services (RDS) gateway provides a gateway service to allow Remote Desktop sessions to access the ADA environment. |
| Analytics Discovery Area (ADA) | Analytics | The Analytics Discovery Area (ADA) is an area of the DAP that allows power users to perform analytics on data stored within the DAP, without requiring the data be transferred or processed on PTV desktop environments. |
| ADA VMs  (IaaS) | Analytics | ADA VMs are a collection of virtual machines, accessible through the RDS Gateway to provide ADA users a desktop environment in which to perform data analytics. The following tools will be installed on these servers; an R programming language implementation, SQL Server tools, and Microsoft Excel 2013.  **IaaS**  While there is a SaaS option for Remote Applications (not desktops), the flexibility of the ADA environment requires that specific VMs be created and managed.  **Scaling Options**  The ADA VMs have been sized appropriately, and will be scaled up if required, or additional VMs can be created and added to the RDS farm.  **Redundancy**  The four ADA VMs provide redundancy. |
| Spark Cluster  (PaaS) | Analytics | Apache Spark is an open source cluster computing framework that provides an in-memory cluster suited to machine learning and big data analytics (as opposed to Hadoop, which is primarily disk based). Spark is provided as part of the Analytics Discovery Area, implemented as HDInsight Spark, Microsoft’s PaaS implementation of Spark.  **Scaling and Redundancy Options**  See the scaling and redundancy options listed in “HDInsight (Hadoop)”. |
| Machine Learning  (SaaS) | Analytics | Azure Machine Learning is a cloud-based machine learning SaaS service, not currently available in the Australian region. It provides scalable machine learning tools with Windows Azure Blob Storage or Azure SQL Database/Data Warehouse data sources to create analytical models.  **SaaS**  As a SaaS solution, PTV need not deploy any infrastructure. The user creates a workspace in Azure ML, and attaches data sources, such as the EDW running on Azure SQL Data Warehouse or Azure Blob Storage.  **Scaling Options**  As a managed service, Azure ML will automatically scale the Machine Learning jobs, increasing the costs based on CPU hours used. |
| Active Directory  (IaaS) | Infrastructure | A requirement of SharePoint 2013 is that it must be joined to a domain. An Active Directory installation is provided within the Azure subscription to meet this requirement.  **IaaS**  At the time of writing, SharePoint IaaS only supports being joined to an on-premises Active Directory. As the DAP cannot extend the CenITex WoVG Active Directory into Azure (due to restrictions by CenITex, including that AD is for whole of government, not PTV-specific), the DAP must create its own internal domain.  **Scaling and Redundancy Options**  The Active Directory IaaS deployment consists of two servers in an availability set, providing redundancy and high availability. If needed, the servers can be scaled up. |
| Windows Azure Storage | Infrastructure | Windows Azure Storage underlies all components listed above. The storage is defined as geo-replicated, storing 3 copies of the data in the Azure Australia Southeast data centre (Melbourne), and 3 copies of the data in Azure Australia East (Sydney). In this way, data is protected from corruption locally, and is available in the Sydney data centre if Melbourne goes offline. |

Estimated consumption (in dollars) and sizing of each Azure component is available in the “DAP Estimated Consumption” attachment at the end of this document.

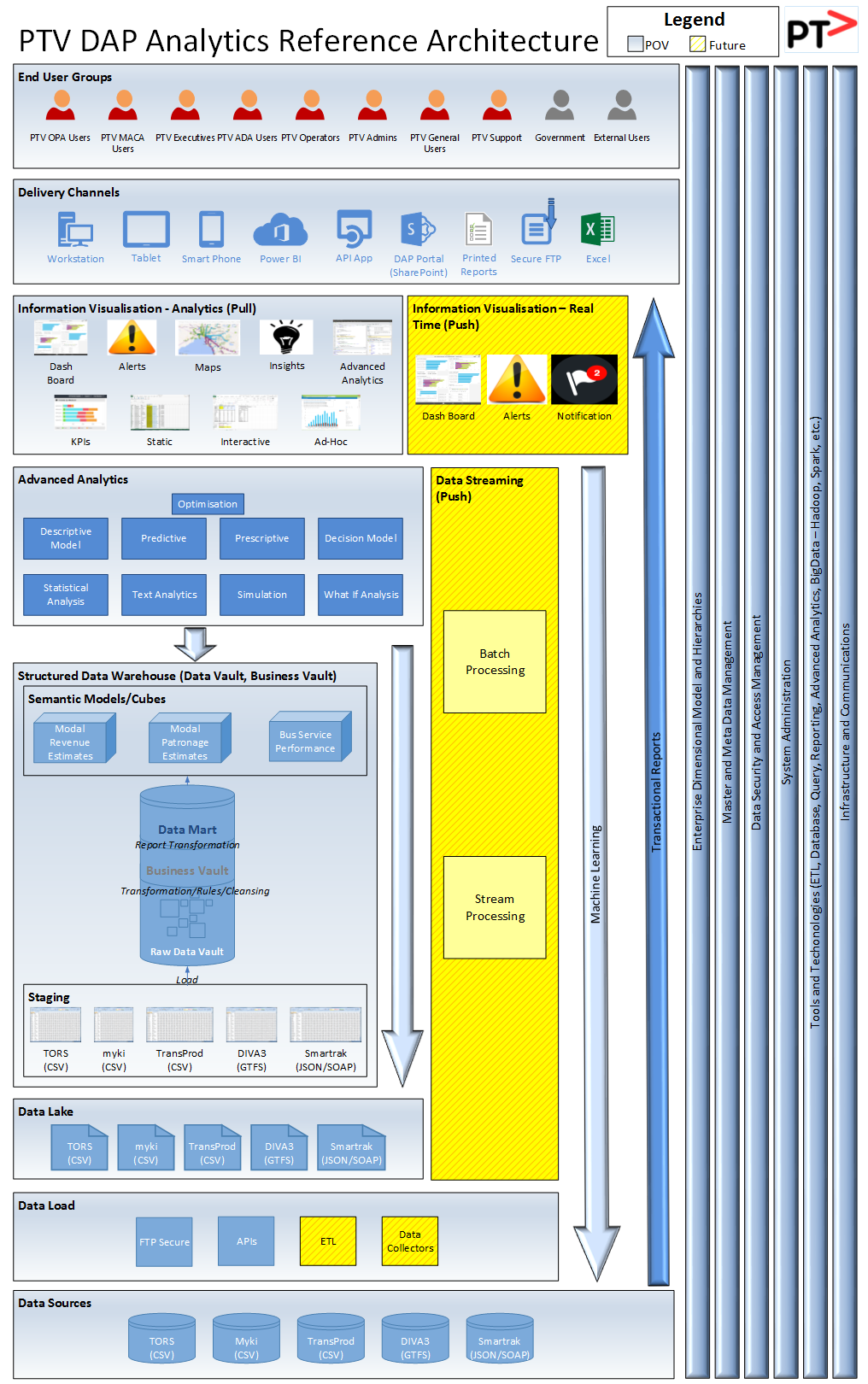
## Migration Plan

The DAP is to be run in parallel to existing systems for a number of months with the following caveats:

* Smartrak Bus Data processed through OPA Small Systems and presented through TI Online will be loaded in parallel, as OPA Small Systems remains the source of data for performance contract management in the short term.
* The process to receive TORS surveys in a CSV format instead of the current emailed Excel spreadsheets will be manually transformed (Excel -> CSV) by the MACA team and uploaded until such time that NTP and HOBAN can provide this interface directly.

## Reference Architecture

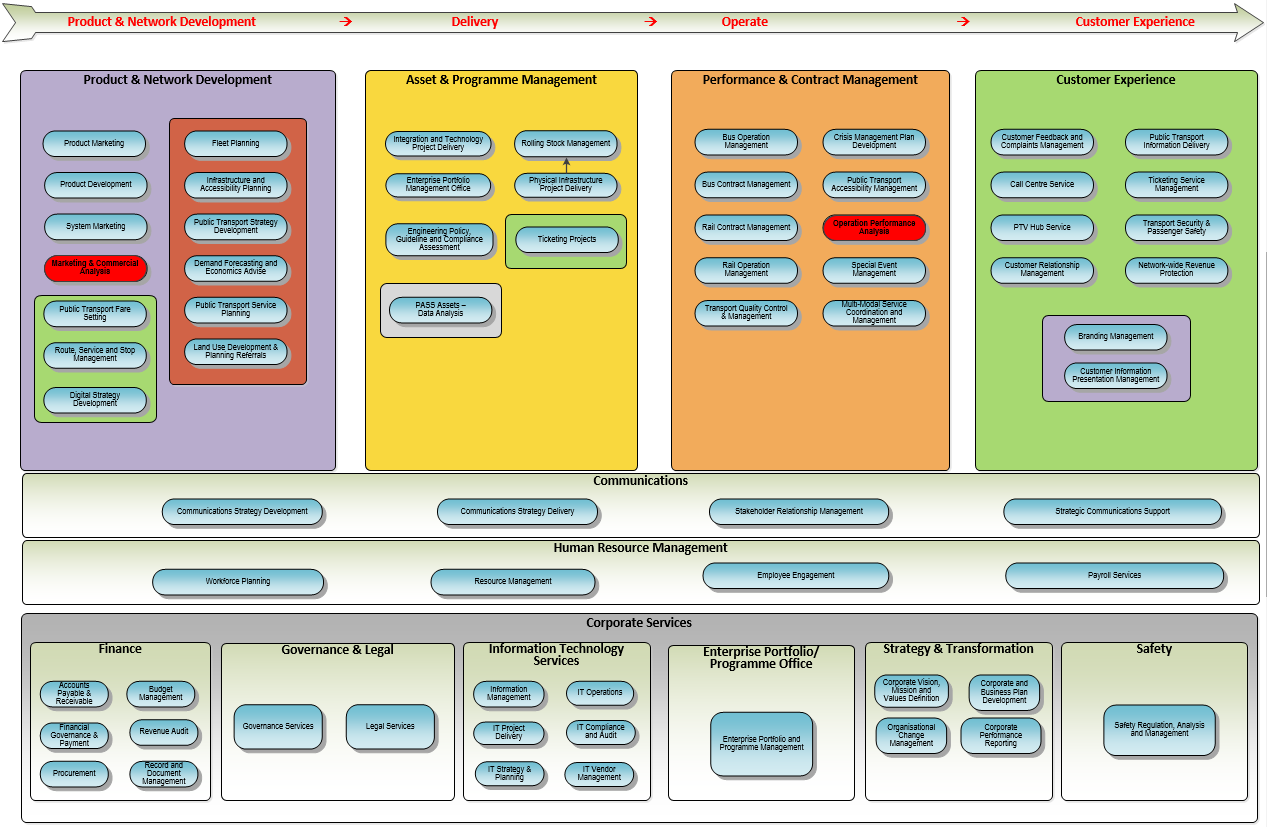
The following diagram depicts the overall reference architecture for the DAP going forward:



# Business Architecture

## Business Functions/Services

In the following diagram, business areas affected by the Proof of Value DAP are shown in red. Note that this diagram reflects the layout of the business prior to the 2015 reorganisation:



### Changed Business Functions/Services

No business functions/services will be changed as part of Proof of Value Data Analytics Platform.

### Decommissioned Business Functions/Services

No business functions/services will be decommissioned as part of Proof of Value Data Analytics Platform.

### New Business Functions/Services

No business functions/services will be created as part of Proof of Value Data Analytics Platform.

## Business Processes

As the Proof of Value Data Analytics Platform is not a direct replacement for an existing platform/solution, changes to existing business processes will be minimal. However, changes will be required and are detailed below.

### Touch On Rate Surveys (TORS)

The Touch on Rate Survey (TORS) data is collected by two external providers, NTP and HOBAN. They currently email these survey results to PTV in Excel Format, where the results are then loaded, cleansed, and validated via a Microsoft Access database maintained by PTV. As part of the Proof of Value Data Analytics Platform NTP and HOBAN and will be required to upload the survey results to the DAP FTPS server, and an automated process will then load and cleanse the data. PTV staff will still be required to validate/sign off the data through an interface presented through SharePoint before the surveys can be used to populate the Patronage Estimate areas of the Business Vault. Details of this process will be described in the SharePoint User Interface Design Document. It is expected that NTP and HOBAN will send one file every couple of weeks. The files will be processed daily as they arrive, but the patronage calculations are performed monthly, after the relevant myki data for the month has been loaded.

#### Current State

Survey data is currently collected by external providers (NTP and HOBAN) and emailed to PTV in Excel format. These results are then loaded into a Microsoft Access database (maintained by PTV), where the data is then (manually) cleansed and validated prior to consumption for reporting purposes.

#### Future State

The survey result files will be uploaded by NTP and HOBAN to the DAP FTPS server, the files will then be loaded into the Enterprise Data Warehouse (EDW). This will negate the need for the current Microsoft Access database, and the manual process to cleanse and validate the TORS data will be automated as part of the EDW Extract Load Transform (ELT) process. PTV staff will still be required to validate/sign off the TORS data before it will made available for enterprise reporting.

A new R algorithm to calculate/estimate patronage has been developed by Symbolix (an external provider). This code has been tested and signed off by PTV; however it has not yet been released to production. As part of the Proof of Value Data Analytics Platform, the output of this algorithm will be loaded into the Enterprise Data Warehouse. The results of this patronage calculation will then be accessed by PTV staff through the various enterprise reporting solutions provided in the Proof of Value Data Analytics Platform.

### Myki Data

#### Current State

PTV staff access Myki data via a Remote desktop connection to an external database server (hosted by NTT).

#### Future State

NTT will create an automated process to extract and upload the files to the DAP FTPS server. The files will then be loaded into the Enterprise Data Warehouse (EDW), as part of the automated Extract Load Transform (ELT) process. PTV staff can then access Myki data through the data marts in the SQL Data Warehouse, or via the Microsoft suite of reporting tools.

### OPA Reporting

No changes to will be required to existing OPA reports until after completion of the proof of value phase.

### MACA Reporting

No changes to will be required to existing MACA reports until after completion of the proof of value phase.

# Data Architecture

## Conceptual Data Entities

The Proof of Value Data Analytics Platform lays the foundation for an enterprise-wide view of PTV’s business, and will incorporate entities from many areas of the business. For proof of value, the following data sets have been identified, sourced from the PTV Enterprise Information Model list of data sets:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Functional Domain | Information Asset | Information Sub-Asset | Source for DAP | System of Entry/System of Record/ Managed Copy |
| Contract Management | Punctuality | Punctuality Bus | Smartrak | MC |
| Management Information | Patronage | Patronage Analysis (by route, stop, time of day etc.) | TORS, GTFS | SoE |
| Management Information | Patronage | Patronage Metro Bus | TORS, GTFS | SoE |
| Management Information | Patronage | Patronage Metro Train | TORS, GTFS | SoE |
| Management Information | Patronage | Patronage Tram | TORS, GTFS | SoE |
| Management Information | Patronage | Touch on Rate Survey | TORS | SoR |
| Operations Monitoring and Control | Stops Operations | Stops Operations Bus | TransProd/GTFS | MC |
| Operations Monitoring and Control | Stops Operations | Stops Operations Train | TransProd/GTFS | MC |
| Operations Monitoring and Control | Stops Operations | Stops Operations Tram | TransProd/GTFS | MC |
| Operations Monitoring and Control | Stops Operations | Stops Operations V/Line Train | TransProd/GTFS | MC |
| Operations Monitoring and Control | Timetable Actual | Timetable Actual Bus | Smartrak | MC |
| Passenger Information | Accessibility at Stops | Accessibility at Stops Bus | TransProd | MC |
| Passenger Information | Accessibility at Stops | Accessibility at Stops Train | TransProd | MC |
| Passenger Information | Accessibility at Stops | Accessibility at Stops Tram | TransProd | MC |
| Passenger Information | Accessibility at Stops | Accessibility at Stops V/Line | TransProd | MC |
| Passenger Information | Accessibility on board | Accessibility on board Bus | TransProd | MC |
| Passenger Information | Accessibility on board | Accessibility on board Train | TransProd | MC |
| Passenger Information | Accessibility on board | Accessibility on board Tram | TransProd | MC |
| Passenger Information | Accessibility on board | Accessibility on board V/Line | TransProd | MC |
| Public Transport Network | Master Data Network | Master Data Network Bus | TransProd/GTFS | MC |
| Public Transport Network | Master Data Network | Master Data Network Train | TransProd/GTFS | MC |
| Public Transport Network | Master Data Network | Master Data Network Tram | TransProd/GTFS | MC |
| Timing Information and Vehicle Scheduling | Route Operational | Route Operations Bus | Smartrak/GTFS | MC |
| Timing Information and Vehicle Scheduling | Route Operational | Route Operations Train | GTFS | MC |
| Timing Information and Vehicle Scheduling | Route Operational | Route Operations Tram | GTFS | MC |
| Public Transport Network | Route Planned | Route Planned Bus | GTFS | MC |
| Public Transport Network | Route Planned | Route Planned Tram | GTFS | MC |
| Timing Information and Vehicle Scheduling | Service Operational | Service Operations Bus | Smartrak/GTFS | MC |
| Timing Information and Vehicle Scheduling | Service Operational | Service Operations Train | GTFS | MC |
| Timing Information and Vehicle Scheduling | Service Operational | Service Operations Tram | GTFS | MC |
| Public Transport Network | Service Planned | Service Planned Bus | GTFS | MC |
| Public Transport Network | Service Planned | Service Planned Train | GTFS | MC |
| Public Transport Network | Service Planned | Service Planned Tram | GTFS | MC |
| Timing Information and Vehicle Scheduling | Calendar Operational | Calendar Holidays | Reference | MC |
| Timing Information and Vehicle Scheduling | Calendar Operational | Calendar School Terms |  | MC |
| Timing Information and Vehicle Scheduling | Calendar Operational | Calendar Special Events |  | MC |
| Timing Information and Vehicle Scheduling | Calendar Operational | Day Type Operational |  | MC |
| Timing Information and Vehicle Scheduling | Calendar Operational | Service Calendar |  | MC |
| Timing Information and Vehicle Scheduling | Timetable Master | Timetable Master Bus |  | MC |
| Timing Information and Vehicle Scheduling | Timetable Master | Timetable Master Train |  | MC |
| Timing Information and Vehicle Scheduling | Timetable Master | Timetable Master Tram |  | MC |
| Timing Information and Vehicle Scheduling | Timetable Master | Timetable Master V/Line |  | MC |
| Timing Information and Vehicle Scheduling | Timetable Operational | Timetable Operational Train | GTFS | MC |
| Timing Information and Vehicle Scheduling | Timetable Operational | Timetable Operational Tram | GTFS | MC |
| Timing Information and Vehicle Scheduling | Timetable Operational | Timetable Operational V/Line | GTFS | MC |
| Ticketing | Ticketing Products | Fare Product Usage | Myki | MC |
| Ticketing | Ticketing Transactions | Scan On/Off | Myki | MC |

The following entities have been identified:

* Network Topology and Timetable
  + Transport Mode
  + Service
  + Service Provider
  + Master Timetable
  + Daily Timetable
  + Actual Timetable
  + Route (Parent / Line / Sub)
  + Stop
  + Service Location
  + Stop Location
  + Vehicle
  + Vehicle Position
* Ticketing – myki
  + Card
  + Fare Product
  + Terminal
  + Terminal Group
  + Zone
  + Scan On/Off Transactions
* Patronage
  + Sample
  + Estimated Patronage

Wherever possible, entities will be named following the TransModel standards for public transport data, and extended if the source systems contain entities that do not match an entity within TransModel. When converting back to a TransModel format, any additional entities will not be shown.

## Logical Data Entities

The following high level entities are defined for the Proof of Value DAP. For more information and greater details on each entity, please refer to the Data Mapping Specification documents.

|  |  |
| --- | --- |
| Service | Function |
| dv.Hub\_AvailableRoute | This Hub entity captures the business key for Available Routes. |
| dv.Hub\_CardStatus | This Hub entity captures the business keys for Card Status |
| dv.Hub\_CardSubtype | This Hub entity captures the business keys for Card Subtype |
| dv.Hub\_Date | This Hub entity captures the business keys for Date |
| dv.Hub\_DepotAllocation | This (reference) Hub entity captures the business key for Depot Allocations. |
| dv.Hub\_Entrance | This (reference) Hub entity captures the business key for Entrance. |
| dv.Hub\_FareProduct | This Hub entity captures the business keys for Fare Product Master. Fare product master lists the actual fare product created on the card. A touch on, touch off, or card transaction interacts with these fare products.  A fare product can be an nHour product (2-hour or more depending on zones travelled), an ePass (variable duration pass measured in days), or a FixedePass (fixed duration pass, e.g Student Pass). |
| dv.Hub\_FareProductStatus | This Hub entity captures the business keys for Fare Product Status which can be associated with each Smart Card |
| dv.Hub\_FareProductType | This Hub entity captures the business keys for Fare Products which can be associated with each Smart Card |
| dv.Hub\_JourneyVehicle | This Hub entity captures the business key for the actual Journey Vehicle (trip) data. |
| dv.Hub\_Location | This Hub entity captures the business key for Location.. |
| dv.Hub\_LocationEntrance | This Hub entity captures the business key for Point of Interest. |
| dv.Hub\_LocationPoint | This Hub entity captures the business key for the actual Point (trip) data. |
| dv.Hub\_Mode | This (reference) Hub entity captures the business key for Mode. |
| dv.Hub\_mykiJourney | This Hub entity captures the business keys for Trip Information as recorded in myki. This is similar to the journey information recorded from Smartrak. The difference is that this data is not system generated and is reliant on the driver entering the correct information. |
| dv.Hub\_mykiRouteStop | This Hub entity captures the business keys for Route Stop as sourced from myki |
| dv.Hub\_Operator | This Hub entity captures the business key for the Operators. |
| dv.Hub\_ParentRoute | This Hub entity captures the business key for Parent Routes. |
| dv.Hub\_PaymentType | This Hub entity captures the business keys for Payment Type |
| dv.Hub\_RemoteEvent | This Hub entity captures the business key for RemoteEvent |
| dv.Hub\_Route | This Hub entity captures the business key for Line (Routes)..  Contains the list of directional travel lines (Eg. Lilydale line to Lilydale, Lilydale line to City) |
| dv.Hub\_ScheduledJourney | This Hub entity captures the business key for Schedule Journey. This is the planned movement of a public transport vehicle from the origin to the destination. |
| dv.Hub\_Service | This Hub entity captures the business key for Service. This is the planned movement of a public transport vehicle from the origin to the destination. |
| dv.Hub\_ServiceException | This Hub entity captures the business key for Service Exceptions. This is the planned movement of a public transport vehicle from the origin to the destination. |
| dv.Hub\_ServiceLocation | This Hub entity captures the business keys for Service Location |
| dv.Hub\_ServiceProvider | This Hub entity captures the business keys for Service Provider |
| dv.Hub\_Shape | This Hub entity captures the business key for Shape. This is the planned movement of a public transport vehicle from the origin to the destination. |
| dv.Hub\_SmartCard | This Hub entity captures the business keys for each Smart Card. This could be the Card Physical Id in the case of an LLSC card or the Card\_Surface\_Id in the case of Disposable card. |
| dv.Hub\_Stop | This Hub entity captures the business key for Stops. |
| dv.Hub\_Survey | This Hub entity captures the business key for Survey files. |
| dv.Hub\_Terminal | This Hub entity captures the business keys for Terminals |
| dv.Hub\_TerminalGroupLocation | This Hub entity captures the business keys for Terminal Group Location |
| dv.Hub\_TerminalStatus | This Hub entity captures the business keys for Terminal Status. |
| dv.Hub\_Time | This Hub entity captures the business keys for Time |
| dv.Hub\_TransactionType | This Hub entity captures the business keys for Transaction Type |
| dv.Hub\_Vehicle | This Hub entity captures the business key for Vehicle. |
| dv.Hub\_Zone | This (reference) Hub entity captures the business key for Zone |
| dv.Link\_CardFinancialTransaction | This link entity captures a financial transaction that occurs. |
| dv.Link\_EntranceTerminal | This link captures the relationship between Entrance and Terminal |
| dv.Link\_FareproductUsage | This link entity captures Fareproduct usage transactions |
| dv.Link\_Journey\_LocationPoint | This link entity captures the relationship between Journey and Location Point. |
| dv.Link\_JourneyVehicleLocationPoint | This link entity captures the relationship between Journey and Vehicle |
| dv.Link\_JourneyVehicleOperator | This link entity captures the relationship between Journey and Vehicle |
| dv.Link\_JourneyVehicleRouteParentRoute | This link entity captures the relationship between Journey, Route and Parent Route. |
| dv.Link\_JourneyVehicleVehicle | This link entity captures the relationship between Journey and Vehicle |
| dv.Link\_LocationPointLocationEntrance | This link entity captures the relationship between a Point of interest and a location. |
| dv.Link\_OperatorRoute | This link entity captures the relationship between Operator and Operator Path |
| dv.Link\_PassengerTrip | This link entity captures the linkage between a Touch on and Touch off transaction with the purpose of identifying a passenger trip |
| dv.Link\_PassengerTripIntermediateStop | This link entity captures the linkage between a Touch on and Touch off transaction with the purpose of identifying a passenger trip |
| dv.Link\_RemoteEventVehicle | This link entity captures the relationship between Remote Event and Vehicle |
| dv.Link\_RouteAvailableRoute | This link entity captures the relationship between Routes and Routes Available |
| dv.Link\_RouteParentRoute | This link entity captures the relationship between Route and Parent Route. |
| dv.Link\_RouteStop | This link entity captures the relationship between Route and Stop. |
| dv.Link\_ServiceLocation\_ServiceProvider | This link entity captures the linkage of a service location to a service provider |
| dv.Link\_ServiceScheduledJourneyStop | This link entity captures the relationship between a Service, Scheduled Journey, and Stops. |
| dv.Link\_ServiceShapeScheduledJourneyRoute | This link entity captures the relationship between a Service, Shape (geospatial points), Route, and Scheduled Journey. |
| dv.Link\_StopMode | This link entity captures the relationship between a Stop of and Mode of transport. |
| dv.Link\_SurveyDetail | This link captures the relationship between Survey, Stop, Vehicle and surveyor SmartCard (Tram and Bus) |
| dv.Link\_SurveyDetailTrain | This link captures the relationship between Survey, Entrance and surveyor SmartCard (Train) |
| dv.Link\_Terminal\_TerminalGroupLocation | This link entity captures the linkage of a terminal to a terminal group |
| dv.Link\_TerminalGroupLocation\_ServiceLocation | This link entity captures the linkage of a terminal group to a service location |
| dv.Link\_TouchOffTransaction | This link entity captures a touch off transaction that occurs. |
| dv.Link\_TouchOnTransaction | This link entity captures a touch on transaction that occurs |
| bv.Link\_SmartrakGTS | This link entity captures the relationship between JourneyVehicle, Stop, Service, Route, and ScheduledJourney. |
| bv.Hub\_Suburb | This Hub entity captures the business key for Suburb. |
| bv.Hub\_Vehicle | This Hub entity captures the business key for Vehicle. |
| bv.Hub\_VehicleType | This Hub entity captures the business key for VehicleType. |
| bv.Link\_ServicePointCleansedServiceCleansed | This link entity captures the relationship between ServicePointCleansed and ServiceCleansed |
| bv.Link\_ServicePointCleansedVehicle | This link entity captures the relationship between ServicePointCleansed and ServiceCleansed |
| bv.Link\_TouchOff\_VehicleJourney | This link entity captures the linkage between a Touch Off and the Journey of the vehicle. |
| bv.Link\_TouchOn\_VehicleJourney | This link entity captures the linkage between a Touch on and the Journey of the vehicle. |



## Data Flow

### Source to Data Lake Data Flow

The DAP relies on source data that comes from many different vendors, each of whom have different configurations and requirements for data extraction and connectivity. For this reason, a standard approach for extracting from source to the Data Lake cannot be provided.

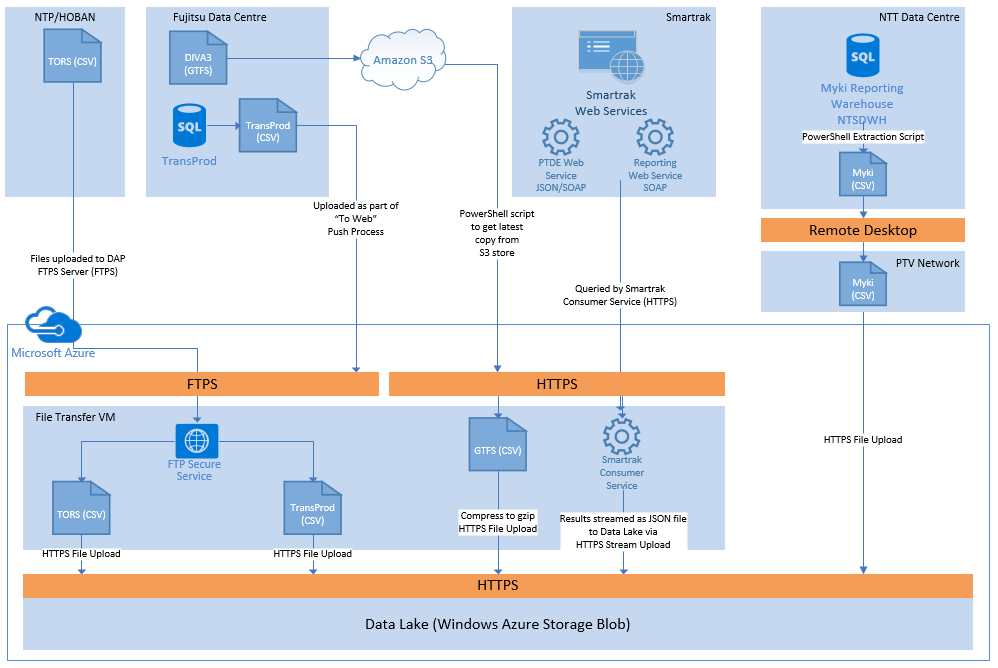


Figure – Data Flow Diagram Source to Data Lake

Each source system is described in section 5.4 – Data Sources

The process to pull data from source systems is standardised wherever possible, with four primary methods for extraction:

* File-based push from a source system to the Azure cloud
  + Untrusted files should be first uploaded to the File Transfer VM to allow a virus scan to take place
    - TORS, TransProd, GTFS are examples
    - Myki data sources extracted by PTV staff and downloaded first to the PTV network or through an ADA VM are considered trusted, and the virus scan will be executed by that machine.
  + During the ETL load process, files are read into the Staging region of the EDW via PolyBase External Tables
  + After loading, source files are archived in the Data Lake
* Direct database access to pull without requiring an intermediate file
  + These data sets are extracted from the source system according to the required filter, and stored in the Staging region of the EDW, whereupon the standard process for loading from Staging to Raw Data Vault occurs.
  + There may be a requirement to persist these data sets in the Data Lake as raw data, extracting from the Staging table into a Data Lake file. This is only necessary if the data set has value in raw format (such as Spark or Hadoop big data analysis that does not want to query the EDW), and the Data Vault modelling technique allows the source system to be recreated at any extraction point in time.
* Query Web APIs, save to Data Lake
  + These data sets are queried through an API, retrieving a data set and saving it to the Incoming directory in the Data Lake.
  + These data sets are generally considered trusted (depending on the data provider), as the query to the data source is querying the data source directly.
  + Smartrak is currently the only example of this.
* Real time data pushed (Future)
  + There are a number of options for processing real time data, including:
    - Azure Event Hubs and Azure Stream Insight
    - Apache Spark (on Azure HDInsight, or an IaaS deployment)
    - Apache Storm (on Azure HDInsight, or an IaaS deployment)
  + As this is a future requirement, the real time landscape available in Azure may be significantly different when the time comes to implement a real time solution. The Data Vault modelling technique provides a model that accepts data in real time as well as in batches, and the real time solution can either process small batches (near-real time), or can insert the data into the Data Vault individually, and flow through into reporting systems (note that SQL Server Analysis Services does not easily do real-time data if it is pre-calculating cubes).

In general, the following hierarchy describes the preference for the data transfer mechanism, in order of most preferred to least preferred:

* Access an available API to pull the data
* Connect directly to a data source if the necessary connectivity requirements can be met
* Push a file to the DAP via FTPS

For Proof of Value, the rationale for the decision for each data source is as follows:

|  |  |  |
| --- | --- | --- |
| Data Source | Method | Rationale |
| Smartrak | API | The Smartrak API is Smartrak’s standard for making data available to downstream users. |
| myki | File Push | No changes can be made to NTT’s environment while the Myki contract is up for renewal. |
| DIVA (GTFS) | File Download | The GTFS feed is published to one location for public consumption, and the DAP must use the same information available to the public for GTFS |
| TransProd | File Push | Fujitsu host the TransProd database and their preference is not to allow direct database access from Azure. |
| Transaction Rate Surveys (TORS) | File Push | NTP and Hoban may not store surveys in a database, and their requirement is to push CSV files to the DAP. |

### Data Lake Onwards Data Flow

This section contains the high level data flow from the Data Lake through to data marts.

The data acquisition process involves extracting, transforming, cleansing and enriching the source data before loading the data marts. The following diagram shows a high level view of the DAP data flow process.

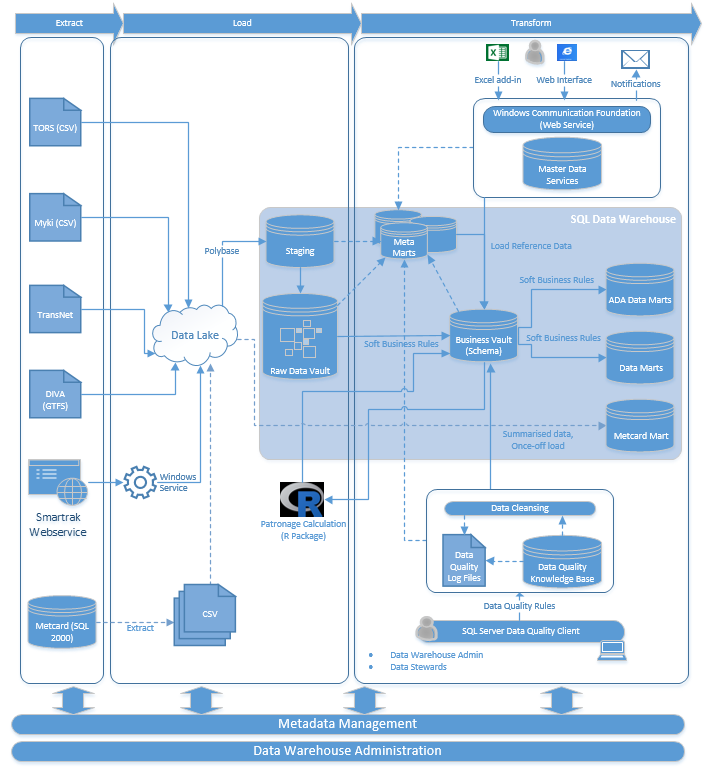


Figure – Data Flow Diagram

Each of these components are described in the following sections.

In the event that a load fails and needs to be reloaded, the ETL process will re-execute. The Data Vault Methodology causes any data that is already listed as the current version of the row will be ignored, so any rows that have successfully entered the Data Vault will be ignored. The Business Vault transformations will then run, and again, any transformed row that is already in the Business Vault is ignored. Finally, the Data Marts are reloaded for the period (existing data for that period is destroyed from the Data Mart, and re-generated from the Data Vault/Business Vault, and the Cubes are reprocessed.

### Abnormal Flow

In a complicated system, many things could go wrong. The table below describes common scenarios and how the DAP will handle these issues

|  |  |
| --- | --- |
| Problem | Handling |
| File corrupted after extraction during transport | With uncompressed CSV files, depending on the corruption, the file may not be successfully read by PolyBase. As the DAP uses GZipped compressed files, any small corruption will result in an unreadable archive, and the ETL job will fail immediately. |
| Data source changes and extracted data is in a different format | PolyBase will be unable to read the source file, and will error out immediately. No data will be loaded, and the job completes with an error, raising an alert. |
| Files fail to arrive on schedule | No data is loaded when the ETL job runs. Operational monitoring will detect that no data has been loaded. |
| Rows fail data validation rules and are not processed into Data Vault. | The job will fail, and raise an alert. This should be a rarity, given how generous the data validation rows are into the Data Vault (data types only) – it should accept all data as part of its audit history. |
| Rows fail data validation rules and are not processed into Business Vault. | Rows are cleansed by data quality rules (detailed in the Data Mapping Specification for each data source), and any row that fails is logged with the rule that it failed. Operational reporting will note the number of rows that failed, but this may not generate an error depending on the severity. Investigation will be necessary, and there may be changes to (and approvals of) reference data in order to support the new data. The BV calculation can then be re-run to continue the data load, superseding the invalid data with new current rows. |
| Rows pass all data validation rules and successfully load into Data Marts but some rows are invalid. | This is the most difficult situation to find, and relies on SMEs checking the data as part of their normal use. If a problem is discovered, new data validation rules or reconciliation checks can be added, and the invalid data will need to be reloaded. If the data is invalid at the source system, changing it there should flow through the system. If it is incorrect in the Business Vault or Data Marts, a suitable data quality rule or reference data will need to be changed, and the period recalculated. The Business Vault will generally add a new row to signify how the value changes over time, while Data Marts will typically be unloaded and recreated for this period (for fact tables) while Type 2 Slowly Changing Dimensions will add a new row. |

As there is no direct source system access, reconciliation can only happen with the Data Lake as a starting point.

Data can be reconciled easily between the Data Lake and the Raw Data Vault, with one of the benefits of Data Vault being it should be possible to reconstruct the source (Data Lake file). All Data Vault loads keep track of the number of rows loaded for each ETL job, and the number of rows staged through PolyBase should match the number of rows loaded into the Data Vault (exception: if the row already exists in the Vault). The number of rows processed for each stage is stored in the Data Acquisition Framework log tables that log all ETL jobs and outcomes (part of the Metrics and Error Vaults).

Reconciling between Data Mart and the source is also possible as most data is day-based. Although the data has been transformed differently, the total number of rows for a specific date can still be grouped with other metrics to provide a reconciled view. The exact definitions of what data sets should be reconciled on a regular basis is undefined, but these can be performed via a stored procedure within the EDW, or through operational reports held in SharePoint.

## Data Sources

This section contains details of the data sources and the extract process for each. This section summarises each data source in business terms, with details on each source in the following sections.

|  |  |  |
| --- | --- | --- |
| Data Source | Description | Processing Requirements and Frequency |
| Smartrak | Smartrak is one of two providers of real time bus tracking data used by Victorian buses, and provides feeds describing how bus services are tracking to the operational timetable, and a log of events that occurs on each bus (such as regular location updates, engine off, doors open, etc). | Downloaded and processed every morning after 4am. Re-processed Lines are processed on demand.  No dependencies on other sources for the Data Vault, but Business Vault requires myki ticketing data to determine bus patronage loading.  Available processing window is <4 hours per day (4am-8am), with processing typically complete with ~1-2 hours to transfer data, 1 hour to load. |
| myki | The myki data source, hosted by NTT contains ticketing data for the myki ticketing system. | The myki Mirror data warehouse is refreshed every Saturday with the most recent backup. DAP will extract data for the previous Saturday-Friday on Monday morning.  No dependencies on other sources.  Estimated processing time is 2 hours to transfer data, and 2 hours to load. |
| DIVA (GTFS) | The GTFS extract from DIVA is an existing open data set provided to the public describing operational timetables. | GTFS is updated manually by PTV. A new file is expected every 4-6 weeks. DAP will check for a new file daily.  No dependencies on other sources.  Estimated processing time is <1 hour. |
| TransProd | The TransProd database stores network topology and timetable information used by the myki ticketing system. It is used by the DAP to enrich the data provided by the GTFS feed | TransProd is updated manually by PTV. A new set of files is expected every 4-12 weeks. DAP will check for a new file daily.  No dependencies on other sources.  Estimated processing time is <1 hour. |
| Transaction Rate Surveys (TORS) | Transaction Rate Surveys are performed by partner companies NTP and HOBAN to survey the number of myki scans at different locations on the public transport network. This data is used by the DAP in conjunction with ticketing data to provide patronage estimations. | Each provider is expected to provide a single file every couple of weeks. New files are detected and loaded daily.  Patronagle calculations cannot begin until the entire month of myki ticketing data has been loaded.  Estimated processing time is <1 hour. |

### Data Volumes

|  |  |  |  |
| --- | --- | --- | --- |
| Data Source | Data Volume (Rows) | Data Volume (MB) | Yearly Growth Rate |
| Smartrak | - | 3 GB per day uncompressed  Approximately 10 TB per year | 30-50% |
| myki | 12 billion rows | 6.21 TB | 2.5-5.6% |
| DIVA (GTFS) | 10 million rows | 800 MB uncompressed snapshots | 1200-52000%, depending on how often snapshots change |
| TransProd | - | < 2 GB total | - |
| Transaction Rate Surveys (TORS) | 512,000 rows per year | - | - |
| Metcard (ATS) Historical Data | 17 billion | 1.5 TB uncompressed | N/A |

### Smartrak

The Smartrak Consumer is a Windows service written in C#.NET that will call the Smartrak JSON web service to obtain the Smartrak data. Any data retrieved will be placed in the Data Lake for processing into the EDW and future analysis.

#### Data Source Description

The Smartrak web services provide SOAP/XML and JSON APIs to query bus data from the Smartrak buses. There are four services currently available:

|  |  |
| --- | --- |
| Service | Function |
| Get Remotes | Gets the list of AVL units available |
| Get Available Lines | Gets the list of Service Lines available |
| Get PTDE for Period by Remote | Gets the PTDE information for a particular remote over a specified period (less than or equal to 24 hours) |
| Get PTDE for Period by Line Name | Gets the PTDE information for a particular Line over a specified period (less than or equal to 24 hours) |
| Get Reprocessed Lines | Gets a list of lines that have been reprocessed by Smartrak |
| Get Depot Allocation Stats | Retrieves a list of all Depots, and how many of each Depot’s routes that have been allocated for a day. |
| Get History (SOAP) | Retrieves history for a specific remote for a day to view all real-time events. Estimated data is 2 MB per day (compressed to 100 KB in Data Lake). |

The web services are documented in the document “PTV Services Interface Design Document: Processed Transport Data Exchange (PTDE)”.

The JSON web service is preferred, and the Data Analytics Platform has been given its own authentication key (not recorded in this document for security purposes).

#### Data Extraction

As Smartrak data may not be finalised until late in the day, the Proof of Value DAP will delay extracting data until 4am, and load the data in nightly batches.

At 4am each morning, a job scheduled by SQL Server Agent will create a new task in the database. Each task records a creation datetime, a started datetime, and a completion datetime to allow for performance metrics to be recorded.

The Smartrak Consumer service maintains a connection to the DAP at all times, and actions are database driven. At 30 second intervals, the Consumer will query the database to see if there are any tasks to do. When a new task is found, the Consumer will create a call to the “Get Remotes” and “Get Available Lines” services. For each Remote and Line returned, the Consumer will create additional tasks in the database to retrieve detailed PTDE data for each Remote and Line, as well as detailed history for each remote (“Get History”). All returned files are stored in in the Data Lake’s Smartrak storage container in the “Incoming” directory.

Once all files have been downloaded (all tasks are completed), the SSIS loading package will run to process the Smartrak data into the EDW.

Additionally, if lines have been reprocessed by Smartrak, the Smartrak Consumer will also periodically call the “Get Reprocessed Lines” service, which will trigger the Smartrak Consumer to re-download and process PTDE data for lines and associated remotes.

#### Data Load

Smartrak is loaded on a per-service per-day level, and the SSIS packages are written at these levels. Therefore, if a single line for a day needs to be reprocessed at any time, that one line’s data can be retrieved and loaded.

#### Future State

As a real-time system, there is potential for the DAP to receive data pushed in real time rather than the batched-pull format described in this document.

There are a number of options for more real-time data:

* The Smartrak Consumer polls the Smartrak web services more frequently (at 1-2 minute intervals), and posts the results into an Azure Event Hub to be processed by Azure Stream Analytics, loading directly into the Raw Data Vault, through the Business Vault and into data marts.
* A change made by Smartrak could post real-time events directly into PTV’s Event Hub to minimise lag time
* PTV can subscribe to a SIRI feed (Service Interface for Real Time Information) provided by Smartrak to provide additional real time data (see <http://www.siri.org.uk>)

Event Hubs and Stream Analytics are Azure components that allow for complex event processing. Stream Analytics performs the complex event processing on each row, and can take immediate action if necessary. Event Hubs and Stream Analytics can process millions of transactions per second, allowing each event to be compared with baselines to detect anomalies, and can power real-time dashboards.

### myki

#### Data Source Description

myki data is currently hosted in NTT's environment, on a server named NTSSQLDWHMIR01. The MACA team have read-only access to this database, named NTSDWH, currently sized at 7 TB.

The MACA team can access the server by using remote Desktop to 10.7.12.10 (a public IP address to NTSSQLDWHMIR01), and logging directly onto the server using a local Windows account. They can then run SQL Server Management Studio to run queries against NTSDWH.

#### Data Extraction

As connectivity is limited to the myki reporting environment, the following interim solution is proposed.

A PTV operator logs onto NTSSQLDWHMIR01 and runs a PowerShell script that takes in a start and end date range:

* The script reads a text file describing all full table dumps to be performed (source table name and target file name)
* The script loops through the text file extracting CSV files from the NTSDWH database and outputting to disk
* The script reads a text file describing all partial datasets to load in (source table name, WHERE filter and target file name)
* The script loops through the text file extracting CSV files from the NTSDWH database and outputting to disk, one file per table per day.
* The script outputs a summary file of what tables were extracted, and timing metrics
* The script compresses (zip) all files in place (50 CSV files = 50 zipped files)

The operator copies these files to their local desktop (1.4 GB estimated) via Remote Desktop shares, and then uploads these files to the Data Lake via a PowerShell script that calls the Azure Blob Upload API, placing them in the myki “Incoming” folder.

**Note:** A detailed extraction process, including source data set descriptions, and timings will be documented in the Data Mapping design documentation.

#### Data Load

myki data is extracted in two different formats. For reference data (small tables), the entire table is loaded into the EDW. For larger transactional tables (scan ons/offs, fare product usage), data is extracted one day per file per table, and loaded independently. Deltas are controlled via the Processing Date.

As PTV only have access to myki data that has been updated weekly, each weekly load will contain seven files for each transactional table. This allows the file to be archived into the correct month in the Data Lake, based on the Processed Date.

#### Future State

It is acknowledged that the proposed method of accessing the data and requiring a manual component is not ideal long term. The future availability of access to the NTSDWH database is currently unknown, but could be one of the following:

* Present a SQL Server endpoint (on port TCP 1433), allowing the DAP’s SQL Server Integration Services to connect and extract data directly from NTSDWH.
* Allow an additional IP address (of one of the ADA machines) to connect to the public IP address of NTSSQLDWHMIR01, bypassing the current requirement to copy data into the CenITex network. The files can be copied to the ADA machine and then uploaded to the Data Lake.
* Run an automated extract of necessary data from within the NTT environment to flat files, and push these files over FTPS or through the Azure Blob Upload API (HTTPS). This extract/upload will need to be scheduled and run from within the NTT network, and appropriate outbound ports (FTPS or HTTPS) will need to be opened.

### DIVA (GTFS)

#### Data Source Description

DIVA provides data in the General Transit Feed Specification (GTFS) format. This provides a customer-centric view of the network topology and timetable data.

#### Data Extraction

PTV currently upload GTFS data to an Amazon S3 bucket, located at <http://s3-ap-southeast-2.amazonaws.com/data.ptv.vic.gov.au/downloads%2Fgtfs.zip> A PowerShell script running on the ETL server will download the file nightly, unzip it, compare the modified date of each file, and, if changed, the file will be placed into the GFTS\incoming directory in the Data Lake for loading into the EDW via SSIS.

#### Data Load

GTFS data is loaded based on each GTFS feed (one per mode). Only changed GTFS files will be loaded.

### TransProd

#### Data Source Description

The TransProd database will supply additional network topology and timetable data not available from the GTFS feeds.

#### Data Extraction

The TransProd database (SQL Server 2012) is managed by Fujitsu, and Fujitsu will be engaged to create the extract process for the data the DAP requires. This will be tied into the “To Web” process that is currently manually started to push changes from the Fujitsu data centre to PTV’s operational web environments. An addition step will be added to upload the extracted TransProd data to the DAP’s Data Lake, into the TransProd\Incoming directory, where they will be loaded into the EDW via SSIS.

#### Data Load

TransProd data is loaded in single batches, with all tables loaded at once. The data is manually released for loading to the Data Lake.

### Transaction Rate Surveys (TORS)

#### Data Source Description

TORS spreadsheets are currently emailed to PTV by National Talent Partners (NTP) and HOBAN. PTV then perform checks and data cleansing and load the data into an Access database.

#### Data Extraction

The DAP project will improve this processing by streamlining the process to load spreadsheets directly into the EDW. The files will be uploaded by NTP and HOBAN to the DAP FTPS server, and a script will move these files into the Data Lake for processing.

#### Data Load

Each TORS survey can be loaded independently into the Data Vault as each survey is a standalone data set. As a new data source is detected in the Incoming directory, the ETL process will stage and load to the Data Vault. Additional approval (presented through SharePoint) and validation steps will be performed by Market Insights prior to the survey being transformed to the Business Vault for use in patronage calculations.

## Enterprise Data Warehouse

Azure SQL Data Warehouse is a turn-key cloud data warehousing and analytics solution. It is based on SQL Server and utilises the massively parallel processing architecture of the Analytics Platform System. It supports features found in SQL server, such as stored procedures, user-defined functions, table partitioning, indexes, collations, and PolyBase (T-SQL front end) to query seamlessly across both relational data in a relational database and non-relational data in common Hadoop formats. It integrates with existing Azure data tools including Power BI for data visualisation, Azure Machine Learning for advanced analytics, Azure Data Factory for data orchestration and movement as well as Azure HDInsight.

### Data Vault Overview

The Data Vault is a relational database. It is the central point of data integration for business intelligence, delivering a common view of enterprise data. “Hard Business” rules (column type standardisation and type validation) are applied during the loading process to the Data Vault. These rules can include normalization and default values (null replacement). The Data Vault in conjunction with the Business Vault will serve as the (data) source for the data marts to service the analysis and reporting services.

Data Warehouses modelled on the Data Vault methodology have the following characteristics:

* **Adaptability**: It adapts to a changing business environment and allows for new data sources to be added without impacting the existing design.
* **Auditability**: Data Vault keeps a comprehensive history and for each record stored in the Data Vault the record source and load date information are captured
* **Performant**: Data Vault is designed to take advantage of Massively Parallel Processing (MPP) style platforms used in Azure SQL Data Warehouse.
* **Integrated**: the data warehouse will contain consistent and integrated data from various sources.
* **Time-variant**: changes to data are tracked and recorded so reports can be produced showing changes over time
* **Non-volatile**: Data in the Data Vault is never over-written or deleted — once committed, the data is static, read-only, and retained for future reporting.

**Note:** The Data Vault model used for the DAP will incorporate common concepts and naming conventions from TransModel to ensure common (industry standard) language is used.

#### Data Vault Core Entities

The Data Vault consists of three core components, the Hub, Link and Satellite.

**Hubs:** The Hub represents a core business concept such as customer, vendor, sale or product. The Hub table is formed around the business key of this concept and is established the first time a new instance of the business key is introduced to the Data Warehouse.

The Hub consists of a business key only, with a sequence ID, load date/time stamp, and record source.

**Links:** The link represents the relationship between two or more business concepts (business keys and is established the first time this association is introduced to the Data Warehouse.

The Link consists of sequence IDs from its related Hubs and Links, a sequence ID, load date/time stamp, and record source.

**Satellites:** The Satellite contains the descriptive information (context) for a business key. The Satellite is keyed by the sequence ID of the hub or link to which it is attached plus the date/time stamp to form a two part key. Note: The Satellite is the only entity in the Data Vault to track change.

The Satellite consists of sequence IDs from its related Hubs and Links, a sequence ID, load date/time stamp and a record source.

These three components are the building blocks of the Data Vault. Together they can be used to represent all integrated data. The Hubs are the business keys, the Links represent relationships, and the Satellites provide the context and changes over time.



#### Data Vault Hybrid Entities

The Data Vault approach has a defined set of hybrid tables used to make the overall deployment more efficient. These are applied on a case-by-case basis as appropriate.

**Point-in-time-table:** The point-in-time table is a modified Satellite that tracks the valid time slices of Satellites surrounding a particular Hub.

**Bridge Table:** The Bridge table is a modified link tables that flattens the relationships between Hubs into a single table for ease of access and performance.

### Raw Data Vault

The Raw Data Vault is the central point of data integration in the EDW. All data in the Raw Data Vault is stored with minimal changes from source data, with the requirement that the source system extract should be able to be regenerated from the Raw Data Vault data. The minimal changes are that data types are aligned (see section 5.5.5.3) and the data is stored in the entity Hub-Link-Satellite format.

### Business Vault

This sub layer (database schema) within the Data Vault represents the data following the application of the soft business rules that may be required to cleanse and validate data, and to apply common transformation rules.

Power users who understand SQL and relational databases can be granted access to the Business Vault. However, the majority of users will not be given access to this layer; they will access data via the data marts.

### View Layer

Between the Business Vault and the Data Mart layers, views are created inside the EDW to simplify the querying of the joining of Business Vault to Data Vault, and to reduce errors introduced when users unfamiliar with the nuances of Data Vault query the Warehouse. These views are then used in direct queries against the EDW, and to assist in populating the Data Marts.

### Database Schemas

The EDW database will contain a number of database schemas. The contents of each of these schemas are detailed in the table below:

|  |  |
| --- | --- |
| **Schema** | **Description/Contents** |
| Staging | Temporary working area to support the process of moving data from various sources to the Raw Data Vault. |
| Raw Data Vault | Staging data is re-modelled into a Data Vault. Acts as the Central point of data integration for business intelligence, delivering a common view of enterprise data. |
| Business Vault | Stores data following the application of Soft business rules. Sub layer within the Raw Data Vault. |
| Metrics Vault | Contains the raw technical metrics for the ETL process and database, including load performance and statistics. |
| Error Vault | Contains the raw technical metrics for errors in the ETL process. |
| Meta Vault | Contains the raw business metadata ontologies/taxonomies/definitions) and physical data model attribute names. |
| Metrics Mart | Data Access layer for the Metric Vault, consisting of views to convert from Vault to Tabular formats. |
| Error Mart | Data Access layer for the Error Vault, consisting of views to convert from Vault to Tabular formats. |
| Meta Mart | Data Access layer for the Meta Vault, consisting of views to convert from Vault to Tabular formats. |
| Data Mart | Data Access layer for the Raw Data Vault and Business Vault. |
| Metcard | Contains the database objects for the legacy Metcard (ATS) database. |
| ADA | An unmanaged region where ADA Power Users can upload data sets and create unmanaged Data Marts. |

More information on these schemas, is provided in the rest of section 5.

#### Staging Schema

The Staging Schema will be used as a temporary working area, where extracted data is stored prior to being loaded into the data vault and data marts. It is a copy of sources primarily utilised for supporting the process of moving data from various sources to the data warehouse. The Staging Area is non-persistent and is emptied prior to the execution of the ETL process. Data will be acquired via a “Pull Mechanism”, delta records will be identified and loaded into the Staging Area using a truncate/insert mechanism, ensuring the ETL process is re-runnable. Transformations are kept to a minimum during the load process in order to make the loads as parallel and independent (from other stage loads) as possible.

#### Staging Entity Structure

The structure of the staging entities will be as close to the source system as possible, with the inclusion of the following fields: Sequence Number, Load Date/Time (recording when the context information is available in the data warehouse), Extract Date, and Record Source (a recording of the source system utilised for data lineage traceability).

Each staging entity is truncated prior to the next load.

For example:



Figure – Sample Staging Table

#### Aligning Data Types

Attributes tend to vary greatly between source systems. To keep the system manageable it is recommend that the source data types be categorised.

The following categorisation will be used:

* Character types such as CHAR or VARCHAR =< 100 will become VARCHAR (100) or equivalent.
* Character types such as CHAR or NVARCHAR >100 and <1000 will become VARCHAR (1000) or equivalent.
* The rest of the character attributes will become VARCHAR (4000) or equivalent.
* Integer values are retained.
* All numeric values will be mapped to a large DECIMAL value. Note: The NUMERIC data type is not supported in the Azure SQL Data Warehouse.
* Floating point numbers are not currently used in PoV data sources, but will be mapped to FLOAT(53).
* DATE and DATETIME types will be mapped to a high precision DATETIME attribute such as DATETIME2.
* Boolean values are converted to bit values True -> 1, False -> 0.

#### Hash Keys

A key feature in the Data Vault 2.0 methodology is the replacement of sequences/identity (integer) attributes with hash keys as the primary key for Hubs, and Links. These hash (surrogate) keys are generated in the Staging schema, as once theses hashes have been generated all other processes can be run in parallel and spread across MPP nodes.

There are three key reasons that Data Vault 2.0 uses hash keys over integer surrogate keys:

1. Not all database engines (relational or non-relational) have the capability or capacity to use natural or business keys for data distribution
2. Not all database engines (relational or non-relational) have the capability to execute EFFICIENT joins on natural or business keys
3. In a Data Vault Model we leverage a many to many relationship table called a Link.  It is made up of multiple keys (from different Hubs).  To join ALL these keys together, would mean replicating the business keys to the Link – resulting in (most cases) a variable length multi-part, multi-data type key set, which would ultimately perform slower than a concisely measured, precise length field.  For Satellites, it means replicating the business keys to each of the Satellites as well.

The DAP will utilise the SHA1 algorithm for all key values, unless a particular data set requires a larger hash. No data sets have yet been identified that would require a larger hash algorithm. SHA1 generates 160-bit keys, with a 1 in 1018 chance of collisions in a dataset with 1.71x1015 key values.

##### Hash Collisions

When utilising hash keys there is a (very) small chance of a hash collision. In simple terms this is when two different data values that go through the mathematical algorithm, and produce the same key value.

There are three available options to address a hash collision:

1. Modify the use of the HASHBYTES() function to use a different algorithm. E.g. upgrade from SHA1 (160-bit) to SHA2\_256 (256-bit) or SHA2\_512 (512-bit).

2. Change the order of the values passed into the hashing algorithm.

3. Change the source data slightly. E.g. change “Hawthorn” to “Hawthorn Station”

Given the small chance of a collision occurring with SHA1 (orders of magnitude less likely than a meteor landing on the data centre), each situation should be considered separately (regenerating all keys for a data set to use a new algorithm vs a slight data modification on one row).

## Data Marts and Cubes

In general Data Marts are the data access layer of the data warehouse. The data in each data mart is tailored for a particular capability or function, and each data mart will be designed in detail with the Subject Matter Experts for that business function. In the DAP solution, Data Marts are used for the following reasons:

* Restructuring the data model: A need for structuring data in a specific data model e.g., restructure to a Dimensional model for OLAP and end user querying/reporting.
* Performance: to offload the data mart to a separate computer for greater efficiency or to obviate the need to manage that workload on the centralized data warehouse.
* Security: to separate an authorized data subset selectively.
* Common dimensions will be placed in a single Data Mart to ensure only one copy of these dimensions exist.
* Data Marts will vary with the granularity of the data being stored. Some will be summarised Data Marts that rely on the Business Vault to provide the transactional granularity, while others will contain all transactional rows inside the Data Mart. At all times, the granularity of the data will be accessible at the transactional level that was sourced from the source system.

A guiding principle of the Data Vault methodology is to enable “Managed Self-Service BI”. The “Managed” part indicates that power users in OPA and MACA can get access to the data stored in the Raw Data Vault or the Business Vault and create their own Data Marts by applying new business rules to transform data into meaningful information. These data marts are stored in the ADA region of the EDW, and there is no dependency on waiting for the Data Warehouse developers to implement a new feature. A single data mart will be created, with views to the following data domains:

* Common Data - Network Topology and Timetables
* Patronage Estimates
* Smartrak Bus Tracking History (PTDE format, aligned to daily/operational timetables)
* Smartrak Bus History (Bus data events, such as engine starting, doors opening, bus location, driver operations)
* Myki Transactions (Fare Product Usage, Scan Ons, Scan Offs)

These domains then feed into the three primary reporting analytics cubes:

* Modal Revenue Estimates
* Modal Patronage Estimates
* Operational Performance - Bus

These cubes are implemented as SQL Server Analysis Services tabular models, with summarised data stored within each cube. DirectQuery is used if needed to drill down to the lowest levels of granularity, preventing the models from being filled with transactional detail, but still granting the ability to query this detail if required. Due to the columnar data store in the Data Marts, DirectQuery is still relatively fast at querying this data.

Each tabular model should be self-sufficient, and querying across models should not often occur. This would usually indicate an incompleteness of the model. If the model is incomplete, joins can be performed in the reporting tools, or by querying the Data Marts directly.

## SQL Server Integration Services

SQL Server Integration Services is the primary controlling application of the Data Warehouse load processes. The Integration Server has SQL Server 2014 installed, and uses SQL Server Agent to schedule the ETL processes. The SSIS Catalog database is installed on this server, and all SSIS packages will be stored in this database.

The following high-level package types will be created:

* Control Packages – coordinates the other package actions listed below.
* Data Lake to Staging – checks that all files for a single load (smallest discrete unit of work) exist in the Data Lake, and loads to the Staging region of the EDW via PolyBase
* Staging to Raw Data Vault – loads Hub structures, then Link structures, then Satellite structures in parallel
* Raw Data Vault to Business Vault – applies business rules and cleans/transforms/enriches data to be stored in the Business Vault tables
* Data Mart Processing
* SSAS Cubes/Tabular model processing
* Archive Files – archives files that have been successfully processed from the Data Lake Incoming folders to Archive.

## Metadata Management

Metadata management is storing information about information in order to provide a consistent, clear and enterprise wide view of information. It includes a set of processes to collect, manage, and deploy Meta data throughout the Enterprise. Metadata managed by these processes includes three categories:

* Technical Metadata describes the physical structures of data and the detailed processes that move and transform data in the environment
* Business Metadata describes the data structures, data elements, business rules, and business usage of data
* Administrative Metadata describes the operation of the data sources including the data warehouse. This includes audit trails, performance metrics, data quality metrics, and other statistical Meta data.

### Technical & Administrative Metadata

The SSIS Catalog will be used to capture the technical and administrative metadata. As part of the SSIS project deployment model, deployed packages can automatically have their logs captured in SQL Server 2014. The four log levels available are as follows: None, Basic, Performance, and Verbose.

The Recommend log level is Basic, as it provides a balance between the level of information captured and the impact to ETL process performance.

Additional detail for each of the log levels is listed below:

|  |  |
| --- | --- |
| **Log Level** | **Description** |
| None | While the None logging level provides slightly better performance than Basic. This logging level does not capture error and status messages, which makes diagnosing problems difficult. |
| Basic | This is the recommended logging level for everyday use. It captures important events (errors and warnings), as well as enough progress information to display what the currently active task is, and how long each task has taken (internally this information is stored in the [catalog].[executable\_statistics] view).  Note: Basic logging does not capture row counts within the data flow. However, this can be achieved using a variety of (other) methods within SSIS. |
| Performance | The Performance log level should be used when you are doing benchmarking and performance tuning for your packages. While it actually logs less messages to the [catalog].[operation\_messages] view than Basic, it captures a lot more events internally to analyse the performance of the data flow components. As a result, there is more overhead during execution – packages run with Basic will run faster than with Performance |
| Verbose | The Verbose logging level captures a lot of events. Verbose should only be used when trying to debug or diagnose package failures. |
| Custom Log Events | The only log level that captures custom log events (i.e. special events that are raised with custom, 3rd party SSIS extensions, or scripts) is Verbose. Due to the overhead that the Verbose level introduces, it is recommended that custom reports are created to capture these custom events. |

The output of SSIS catalog logging will be loaded into the Metrics and Error Marts, where selected users will be granted access to view metadata reports published to SharePoint and the data can be accessed by internal processes. The type of information captured in the Metric and Error Marts is detailed below:

#### Metrics Mart

The Metrics Mart will be used to store the technical metrics for the:

* Load Process: Run-times, completion rates, insert numbers, update numbers, and row counts.
* Database: Growth of tables, files, and indexes. Information about CPU utilization, RAM access, I/O throughput and I/O wait times.

#### Error Mart

The Error Mart is used to store the technical metrics about errors in the loading process and other functions of the data warehouse. The error mart can also be used to store poor quality data, or data that breaks business rules, increasing the visibility of bad data.

### Business Metadata

The business metadata will be maintained and managed by Master Data Services (MDS) and stored in the Meta Mart. MDS enables business owners/data stewards to maintain core business entities. Ensuring uniform, accurate, and semantically consistent business definitions are published for analysis and reporting, while ensuring IT process control, security, and auditability is maintained.

#### Meta Mart

The Meta Mart is used to store business metadata (ontologies/taxonomies/definitions) and physical data model attribute names. Also stored are Functions (for translation) and technically implemented business rules that the ELT process will follow to interpret the data.

**Note:** SQL Server Integration Services (SSIS) will be used to extract and load data from the Master Data Services (MDS) database to the Meta Mart. Once loaded to the Meta Mart, the data can be used across all reporting areas, to assist with dynamic reporting requirements (dataset owners, disclaimers, etc).

The following diagram provides a high level view of the metadata capture points within the ETL process flow:



The ETL audit checkpoints highlight the metadata capture points for information not captured in SSIS catalog logging. This can include row counts (dependent on level of logging selected) and records inserted, updated, deleted, discarded, or rejected.

## Master Data Services

The MDS solution provides the components to centralise the storage and maintenance of the master data. It provides a configurable database (SQL Server database), user interfaces (Excel Add-in and a web UI), a configurable business rule engine, a mechanism to extract the catalogued master data for use in downstream systems and a full security model.

The Master Data Services infrastructure (Web UI and Database) are hosted on the Integration Services server.

**Note:** The MDS database will only be accessible from the ADA environment. Therefore, those responsible for maintaining master data will require access to the ADA environment.

The following diagram shows a high level view of the MDS architecture:



The DAP primarily stores **managed copies of reference data**, and is not the source system for Master Data. The DAP does not make a good MDM solution in and of itself as it has no way to write-back to source systems with quality rows. The only exception is the Access database currently being used to assist in TORS survey processing, which will be converted from Access to a SQL Server tables and stored on the Integration Server, with the current data sets being loaded into the Data Vault nightly. As all reference data is stored in the Vault format, full lineage and history of this reference data is stored, allowing it to be queried as it was at any point in time in the past.

## Data Quality and Transformation Process

The datasets fed into the EDW, play an important role in PTVs operational, tactical, strategic, and financial decisions, and must therefore be accurate. However, there is a risk of data quality issues with these datasets that require a quality review and transformation process, involving a review and adjustment of the data that is received from the operational systems.

The DAP architecture aims to provide the flexibility that PTV requires without compromising Data Warehouse best practices and at the same time providing a mechanism to improve the quality of data overtime.

The following diagram shows the data flow and quality process for incoming datasets:



The DAP implementation will use DQS to perform data cleansing. DQS is a knowledge-driven data quality product. It enables administrators and data stewards to build a knowledge base of data quality rules which is then used to perform a variety of critical data quality tasks, including correction, enrichment, standardization, and de-duplication of your data.

DQS consists of Data Quality Server and Data Quality Client, both of which are installed as part of Microsoft SQL Server 2014. Data Quality Client is a SQL Server shared feature that administrators and data stewards can use to perform computer-assisted data quality analyses and manage their data quality interactively as well as defining data quality rules in the DQS knowledge base.

The DAP solution will use the DQS cleansing components from within ETL components to perform data quality processes according to rules that are defined by data stewards in the DQS knowledge base. Best practices for data cleansing are:

* Data must be cleansed in the source system
* The cleansing process must record all data quality issues
* Data can be changed in the Business Vault as part of the cleansing process but these changes must be limited to changes that would not change the nature and meaning of the underlying data. Below are some examples:
  + Reformatting a date - Acceptable
  + Change of Address - Not acceptable
  + Reformatting an address – Acceptable
  + Populating data with derived data – Acceptable.

## Downstream Uses

While the Proof of Value is not currently scoped to deliver external feeds to downstream consumers, there are a number of ways this can be achieved programmatically, including:

* WCF OData Services, hosted as an Azure Service
* ASP.NET Web API application, outputting JSON
* SSIS jobs to extract from the EDW to a file, hosted on the FTPS server, or sent to a remote server

ADA users can also create these external feeds to run within the ADA environment, or they can be configured to run on production hosted on either the SharePoint servers, the Data Transfer host, or an additional web server.

## Future Consideration

The DAP solution has been designed to allow for flexibility and scalability, and utilises the power of the Microsoft BI stack to provide ongoing new features. One area where the Proof of Value DAP is currently lightly defined is real-time analytics.

As mentioned in section 5.4.2.4, Real-Time data can be integrated into the DAP by use of Azure Event Hubs and Stream Analytics. Alternatively, if Apache Storm is a better fit for the real-time computation requirements, a Storm implementation can be created (either on IaaS virtual machines, or with HDInsight Storm, which is in preview at time of writing).

### Scalability

Functionality scaling can be achieved by adding additional data sources, following the same process of Extracting to Data Lake, then loading into the EDW and then into Data Marts. Due to the Data Vault modelling methodology, additional data sources can be added to the Raw Data Vault without impacting any existing data sources, and transformed into the Business Vault. Once in the Business Vault, new data sources can be transformed into new Data Marts. An existing Data Mart that would receive new functionality from the additional data source will not be touched – instead a new Data Mart will be created, and once all reporting functionality has been transferred to the new Data Mart, the original Data Mart can be decommissioned.

## Data Migration Plan

### Metcard Data

The Metcard (Automatic Ticketing System, or ATS) dataset is the only data set to be migrated from the existing source system to the Data Analytics Platform.

#### Current State

The ATS dataset is currently hosted in a SQL Server 2000 database on server p00823, in the CenITex Burwood data centre. Connectivity is available directly from PTV desktops to the SQL Server endpoint (TCP 1433) through standard SQL Server tools. The estimated size of this data is 2.5 TB uncompressed.

All large volume data sets are partitioned into separate monthly tables.

There are three main database areas:

* ATS\_REF\_01 - containing the raw transactional and reference data. This is an Operational Data Store database.
* ATS\_RPT\_01 - containing the data cubes and associated dimensions. The cube summary tables will be loaded into the DAP Metcard Data Mart as is. The Proof of Value DAP will not create an SSAS cube on top of these summary tables, but PTV may perform this activity.
* ATS\_QRY\_01- a working space for users to store result sets and other user created or imported data sets. This also contains the tables underpinning the Metcard-based Station Patronage Model.

The Metcard dataset is not in daily use; it is only used occasionally to answer specific ad hoc questions. The export process will not be disrupting any business processes.

#### Target State

The ATS dataset can be loaded into three possible areas:

* The Data Lake, stored as flat files in blob storage
* The DAP Azure SQL Data Warehouse
* A separate Azure SQL Data Warehouse to remove the bulk of raw transactional data from the DAP data warehouse.

All source files extracted from the ATS system will be stored in the Data Lake as is, regardless of whether they are also stored in a relational database.

Due to the bulk of the ATS\_REF\_01 transactional data, this will be stored in the Data Lake only, to be accessed via PolyBase if required. The server used to migrate the ATS data will still be available (turned off unless needed), and can be used to view the data in SQL Server 2014 format.

The ATS\_RPT01 dataset containing summarised cubes and dimensions and the Metcard Station Patronage Model will be stored in the Metcard Data Mart. This will be stored as is, with only minor changes where required if it uses data types that are not supported by the SQL Data Warehouse.

The p00823 server that currently hosts the historical ATS data can be decommissioned once the data is safely in the DAP, however this activity is not in scope of the DAP.

A future use of the Metcard data may be to translate it into a unified format to enable reporting over the Metcard and myki data using the same format. As the Metcard data has a much larger grain than myki, this unification exercise is out of scope for the POV, but the data is provided in the EDW for analysis in the ADA environment.

#### Migration Process

The process for migrating the ATS dataset to the DAP environment is:

* Identify required tables to be extracted.
* CenITex provide compressed backups of the ATS databases
* Provision a SQL Server 2008 R2 and 2014 instance in the DAP Azure environment to restore the databases. Upgrade all databases to SQL Server 2014 format (using SQL Server 2008 R2 as a go-between, as there is no direct upgrade path from SQL Server 2000 to 2014).
* Extract each table to CSV format and compress the file using gzip. Large datasets that are partitioned by time period will be extracted in monthly chunks (these should already be in the format of one table per month, due to SQL Server 2000 not having table partitioning functionality).
* Upload all files to the Data Lake via Azure Blob Upload API (HTTPS). The files will be located in the Metcard container of the MACA storage account.
* Bulk load ATS\_RPT\_01 and ATS\_QRY\_01 data sets into the Metcard Data Mart via PolyBase (PolyBase can access zipped files without decompressing first).

### Bus Tracking System Data

All historical Smartrak data available from Smartrak will be loaded using the same process as loading new data, in that each day is extracted by the Smartrak Consumer and loaded into the EDW.

Historical BTS data (non-Smartrak) that can be easily converted to the same format as Smartrak data held in the Data Mart will also be loaded.

### myki Ticketing Data

The full historical view of myki data will be loaded into the EDW. The process for this is the same as the process for loading new data, in that each day will be extracted to a compressed flat file and processed by the DAP. As there is approximately 1000 days of back-processing, this data will be extracted and loaded in a staged process, one month at a time.

# Information Architecture

## Reports and Formats

The following reporting styles will exist in the Data Analytics Platform. While it appears that there are multiple options for each report type, in general, static, predefined reports will be created and presented through SQL Server Reporting Services, and self-service BI reports will be created in Power BI. Note there are no pre-defined reports to be built as part of the Power BI deployment – all use of Power BI is user-generated. All developer-produced reports in the Proof of Value DAP are SSRS static reports and dash boards presented through SharePoint.

The reason for multiple reporting options is that no one tool in the Microsoft BI stack provides all reporting options. Power BI is the clear direction that Microsoft are headed for self-service BI, but Power BI is still in active development, and while features are being added monthly, some reporting tasks are better performed in other tools.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Report Usage | Reporting Platform | Device | Developer | Notes |
| Static Reports/ Dashboards | SQL Server Reporting Services (SharePoint) | Desktop, Mobile | BI Developer/ Reporting Power User | Static reports and dashboards that allow little-to-no user customisation are presented through SharePoint. |
| Pre-Defined Reports | SQL Server Reporting Services (SharePoint) | Desktop, Mobile | BI Developer/ Reporting Power User | Complex pre-defined reports that require a high level of customisation or specific formatting requirements not possible through other tools. |
| Operational and Security Reporting | SQL Server Reporting Services (SharePoint) | Desktop, Mobile | BI Developer/ Reporting Power User | Operational reports showing the state of the DAP data processing jobs, as well as usage statistics on each report/SharePoint page, and report generation timings. |
| Ad hoc, Self-Service BI | Power View Silverlight (SharePoint) | Desktop | Anyone | Self-Service BI for internal users is preferred by Power BI, but Silverlight is still available as an option, particularly for External users. |
| Ad hoc, Self-Service BI | Excel Services (SharePoint) | Desktop | Anyone | Self-Service BI for internal users is preferred by Power BI, but Excel Services is still available as an option, particularly for External users. |
| Ad hoc, Self-Service BI | Excel 2013 Power View | Desktop | Anyone | Divorced from Silverlight and SharePoint, but requires a direct connection to the data source, limiting usage to the ADA users in the short term. |
| Ad hoc, Self-Service BI | Power BI | Desktop, Mobile | Anyone | User-defined reports, shareable within the PTV organisation. |
| Mobile Dashboards | Power BI | Desktop, Mobile | Anyone | User-defined dashboards by selecting tiles from multiple Power BI reports |
| Analytics Discovery Reports | ADA- Based Reporting | Desktop | ADA Users | User defined reports to assist ADA users in interpreting data analytics |

Report templates and design are specified in the Report Specification Master Document.

In general, reports are presented through SharePoint by functional area, with each functional area having a separate SharePoint site that can be granted separate permissions. These reporting areas are:

* Patronage
* Revenue
* Operational Performance

SharePoint provides usage and performance analytics of each report, allowing the site owners to review how often each report is being used and what the quality of user experience is like.

### SQL Server Reporting Services

Presented through the SharePoint portal, SQL Server Reporting Services reports are best used for complicated reporting uses and dashboards. These reports will be those whose needs are clearly defined, and each report fills a specific business need. SSRS reports are the most capable of all, and support complex reporting requirements that the other report types do not.

As SSRS reports are presented through SharePoint, they can be viewed through the web browser of mobile devices.

### Power View

Power View reports presented through SharePoint 2013 are a Silverlight application (Windows desktop only – not mobile) that allows users to perform self-service BI queries.

SharePoint acts as a gateway to allow PTV users to query the SQL Server Analysis Services semantic model without requiring connectivity directly to SSAS.

The Silverlight application is no longer under active development, with Power BI and Excel 2013 the primary way forward for Power View-type queries.

### Excel Services

Presented through SharePoint, Excel documents can be uploaded to SharePoint to share. As no data is stored in the Excel workbooks, a direct connection to the data sources is required, so Excel services will primarily be used from the ADA area, limiting Excel Services only to those users with ADA access. When Excel 2013 is deployed across PTV, the limitation of only ADA users having this ability should be removed, using SSAS as a data source presented through SharePoint.

### Power BI

Power BI reports are the primary reporting tool for self-service BI and sharing within PTV, as Power BI is organisation-based. External organisations require Power BI Pro licenses, and need a Power BI gateway to be configured in their Power BI portal to allow connectivity to the DAP data sources. SSRS and Power View reports do not have this limitation, and will prove to be an easier way of providing reports to these user groups.

Power BI is under heavy development from Microsoft, and new features are being added monthly, and sharing with external parties is the most requested feature at this time.

Power BI provides an excellent dashboard interface on mobile, without needing to log into the SharePoint portal.

### ADA Based Reporting

The Analytics Discovery Area does not have specific reports as such, but contains an installation of Microsoft Excel 2013 (providing the latest Microsoft BI functionality), and has direct connectivity to the DAP data sources. Additionally, Power BI Desktop will be available in ADA to design reports.

ADA reporting can be considered personal reporting – any reports generated will not be shared to a wider audience directly.

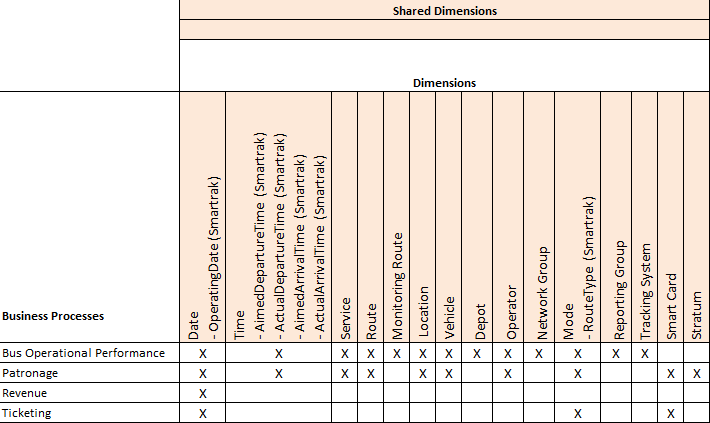
## Dimensional Matrix

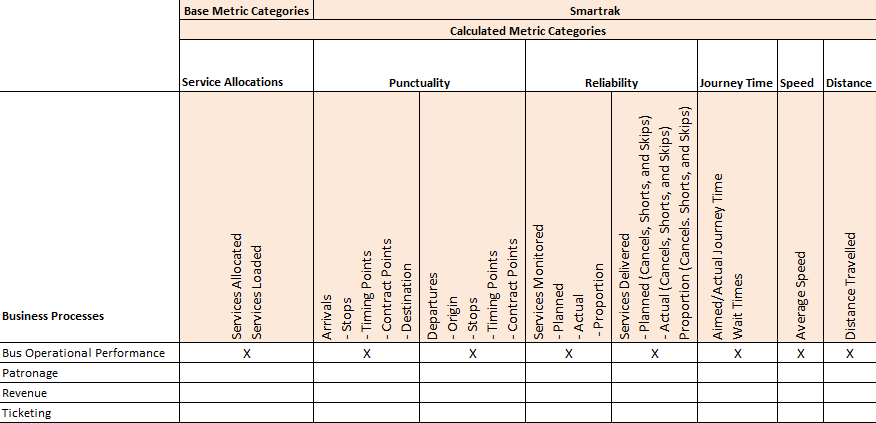
The DAP transforms the Business Vault into a Dimensional structure for reporting purposes. As the EDW is a massively parallel data warehouse, the reporting data mart need not be divorced from the EDW itself, reducing the need for multiple identical copies of dimensions across many data marts.

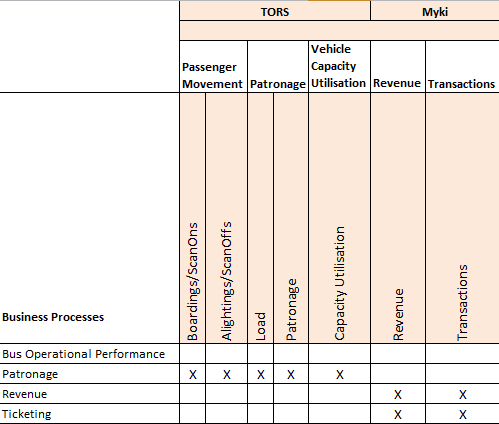
The Data Vault is modelled from an Enterprise-wide view, with each Entity in the Data Vault referring to a single type of entity, effectively conforming entities (and thus dimensions) at the point of the Data Vault, and flowing upwards through to the Business Vault, and into the Data Marts.

Myki data will appear in a similar dimensional format to how it is available in the NTT data warehouse, given that warehouse is the source system for the DAP. It will have additional data produced from the application of PTV Ticketing business rules.

The Bus Matrix is embedded in section 13.3 Attachments, but the content is copied below for reference:







## Data Dictionary and Business Rules

Due to the size and complexity of the Data Dictionary and Business Rules, these are documented in the Data Mapping Specification documents.

# Application Architecture

## Data Lake - HDInsight

A Data Lake is a single repository for all information for an organisation. Data has value that may not be apparent today, and traditional data management has restrictions such as needing pre-definition of schema, the cost of storing large data sets, and the propagation of different data silos.

The Data Lake allows this data to be stored in an unstructured fashion prior to requirements or schema being defined, and can be accessed by HDInsight and PolyBase for operational and exploratory analysis.

While Microsoft has announced an Azure Data Lake product that integrates with Active Directory for security, auditing and performance management, there is no set release date. Azure Data Lake is built on HDInsight technology, so the DAP Data Lake will likewise be built on HDInsight.

This consists of a 4-node HDInsight cluster with Windows Azure Blob storage. HDInsight uses a HDFS API to appear like a HDFS filesystem, but the files are stored on Azure Blob Storage, preventing the need for individual disks to be attached to each node of the HDInsight cluster to store files, and permitting practically unlimited growth without managing individual nodes. The Azure Blob Storage is made up of a number of different storage accounts and containers to allow controlled access to the data stored in the Data Lake.

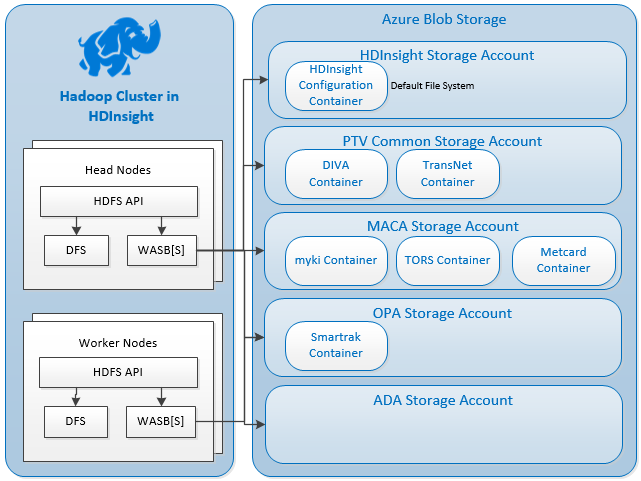


Figure – HDInsight Storage Architecture

As with all Azure services, storage is a pay-what-you-consume model, without concern to hitting storage limits.

As the role of a Data Lake is to be a main repository of data for the solution, it needs to be accessed by all analytical components of the DAP. All incoming source data will first be stored into the Data Lake, and from there processed by SQL Server Integration Services and relevant data will be stored in the Azure SQL Data Warehouse. In future, Azure Data Factory can also access the Data Lake to process data prior to perform analytical tasks.

Data will not be removed from the Data Lake – the intention is that the Data Lake will store all data history for as long as PTV deems the data is required and has value. If hot-access (in the EDW database) to historical data is no longer required, data can be unloaded from the EDW and retained within the Data Lake. No data will be deleted from the system.

Data from multiple sources, including the ADA environment and the Azure SQL Data Warehouse can be extracted and placed into the Data Lake for others to consume.

As the Data Lake holds sensitive data that must be restricted to a sub-set of users, the Data Lake will be partitioned according to access levels, and only those users with the proper keys will be able to access that data.

For operational processes, PolyBase will stage data from the Data Lake through the Windows Azure Storage Blob interface, bypassing the HDFS API, and not requiring compute power from the HDInsight cluster. This allows the compute to be reserved for ADA purposes.

### Data Lake Storage Accounts and Containers

There will be a storage account dedicated for each business unit. The ADA will also have a dedicated storage account. There will be one PTV Common storage account, for data common to all business units.

Each subject area/data source will have a dedicated storage Container. The use of storage accounts and containers is used primarily as a security mechanism, and secondarily as a performance mechanism (storage accounts grant 20,000 IOPS per storage account). Users that have access to a Storage Account have access to all containers and files stored in that Storage Account, and Shared Access Signatures are keys (implemented by globally unique IDs) that provide time-limited access to containers (either read-only or read-write).

Files stored in the non-ADA storage accounts are classified as integral to the DAP/EDW load, and write access is only provided to the DAP. Only the ADA area allows for read-write access. While it is possible to grant read/write access to a new container in the non-ADA storage accounts, this is discouraged to separate ad-hoc use from the Data Warehouse system. Additional storage accounts and containers can be created to cater for any new requirements.

Jobs running inside HDInsight can specify the location of the files to be processed, and the SAS key required to access those files.

Data in Azure Blob Storage is stored as a key/value pair, and there is no directory hierarchy. To simulate a directory structure, the "/" character can be used within the key name to make it appear as if a file is stored within a directory. Files stored in the Data Lake will be stored in each container under simulated “directories” according to the year and month.

To separate incoming files from files that have been “processed” and uploaded to the Enterprise Data Warehouse, incoming files will be placed under an “Incoming” folder. Once processed, the data files will be moved out of the incoming folder into the permanent location in the Data Lake.

For data that will not be loaded into the EDW, the processing will be a move to the permanent location.

The resource URI would then be of the structure:

https://<storageaccount>.blob.core.windows.net/<container>/<year>/<month>/<filename>

For incoming files, the structure is:

https://<storageaccount>.blob.core.windows.net/<container>/incoming/<filename>

## SQL Server Analysis Services – Semantic Layer

Analysis Services provides a range of solutions for building and deploying analytical databases used for decision support in Power BI, Excel, Reporting Services (SSRS), and other business intelligence applications. The SSAS is used in the DAP solution to build models of data in the data warehouse such as cubes which can be queried, analysed and compared using various reporting tools such as Power View, Power BI, SSRS and Excel.

The basis of any Analysis Services solution is a Business Intelligence Semantic Model (BISM) and a server instance that instantiates, processes, queries, and manages objects in that model. BISM is a metadata layer that describes entities and relations between them in business user oriented manner. It is built by the development team and will be used by all the reporting tools. It allows the SSAS to reach a much broader user base.



Figure – SQL Server 2014 Analysis Services supported BISM models

The DAP solution will mainly use a tabular model as it can be used by a much broader user base. The SSAS Tabular Model is an In-Memory database that utilises the xVelocity in-memory analytics engine to process and compress data. This in-memory columnar storage engine has been optimized for high performance analysis and exploration of data. Tabular has the added advantage of being much easier to understand and build than the multidimensional model.

However the DAP solution may include multidimensional models side-by-side with the tabular model if required. In that case a new instance of the SSAS must be installed on the same server or another server and configured for the multidimensional model. This will be determined during Data Mapping and Reports Specification development.

## SQL Server Analysis Services Server Infrastructure

The Analysis Server can scale out if required by addition of more servers to the Load Balanced Analysis Servers. This is useful as the number of cubes and analytical reporting activities increases. However initial configuration is based on a single SSAS server with 2 SSAS instance (1 with dimensional model and 1 with tabular model).

### Disk Storage Requirements

The Analysis Server stores cubes and In-memory cache data. Size of cubes and in-memory cache depend on the size of the fact tables and dimension members.

During processing, SSAS stores copies of the objects in the processing transaction on disk until it is finished, and then the processed copies of the objects replace the original objects. Therefore, sufficient additional disk storage must be provided for a second copy of each object. For example, if a whole cube will be processed in a single transaction; sufficient hard disk space to store a second copy of the whole cube will be required.

Disk volumes can be added easily to Azure Virtual Machines as required, and exact sizing will be determined during the development phase.

### Memory and Processor

Analysis Services needs more memory and processor resources in the following cases:

* When processing large or complex cubes. These require more memory and processor resources than small or simple cubes.
* When the number of cubes within a single database increases.
* When the number of databases within a single instance of Analysis Services increases.
* When the number of instances of Analysis Services on a single computer increases.
* When the number of users who are accessing Analysis Services resources simultaneously increases.

The Microsoft product group recommends 4 to 8 GB of memory per processor core, but this can be more if queries return very large result sets.

For initial phases of the DAP implementation a single 64-bit CPU with 4 cores and 16GB RAM would be sufficient for initial phase. The server performance can be monitored in production and if required it can scale out or up.

## Analytics Discovery Area

The Analytics Discovery Area (ADA) is used to provide an environment where PTV’s data analysts and data scientists can run analytical queries to provide new insights on PTV’s data.

Estimating the computational requirements of the ADA area is difficult ahead of time, but the DAP has been costed to allow all components of the ADA environment to run 24x7. Storage costs are $0.06 per GB per month, providing 6 copies of the data (3 copies in the Australia Southeast data centre, and 3 copies in Australia East). There are no practical maximum limits of storage (500 TB per storage account, 100 storage accounts per subscription, and additional subscriptions can be added and the storage used by the primary DAP subscription).

Additional discussions held in December 2015 have noted that there is no real need to provision a Development and Test ADA environment for processing queries, and so the money saved by not requiring these environments can be used on an ad hoc basic to scale the ADA environment. For example, the Spark cluster can be doubled in size when necessary. While costs vary depending on the size of the additional virtual machines, a good estimate is $0.50 cents per hour per additional virtual machine, and HDInsight (Hadoop and Spark) clusters can be scaled on demand. Alternatively, a temporary Spark cluster can be spun up on demand.

If users in the ADA environment produce a calculation or model that is deemed able to go into production there are a number of options:

* If the result is a once-off dataset that is useful to be included in reports, ADA users can use SQL Server Data Tools to upload the dataset into the ADA section of the EDW. Reports can then join data from existing data marts to the dataset in the ADA environment.
* If an ADA user creates a new data mart based off existing data using either T-SQL stored procedures or SQL Server Integration Services, the process can be deployed and scheduled through the standard data warehouse load process. CGI can assist in ensuring the process fits within the DAP ETL framework.
* Spark machine learning algorithms and models can be initially tested and implemented in a production fashion on the ADA Spark cluster. If warranted, an additional Spark cluster can be provisioned solely for production use.

Exact processes for requesting and implementing scaling ADA and deploying additional features to the main EDW will be developed over the first year of use. The DAP project team will (during development and ongoing) examine Azure consumption monthly and report on areas of over-consumption.

### Virtual Machines

Users will connect through an RDS gateway (https) to run either a standalone application (such as Excel or SQL Server Management Studio), or to get a virtual desktop on one of the four ADA virtual machines. These VMs run Windows, and have the following tools installed:

* SQL Server Management Studio
* SQL Server Data Tools (Azure SQL Data Warehouse connectivity)
* SQL Server Data Tools – BI (SSRS/SSIS/SSAS authoring)
* R
* Python
* Excel 2013
* Power BI Desktop

Other tools can be installed by request.

### HDInsight (Hadoop)

The original design utilised a separate HDInsight cluster for the ADA area to keep the Data Lake unaffected by ADA queries.

Operationally, the Data Warehouse queries data in the Data Lake via PolyBase direct to Windows Azure Storage Blobs (WASB), and no compute power is required. This leaves the Data Lake HDInsight cluster’s compute power available for running Hadoop queries.

### Azure SQL Database (Sandpit)

A requirement is to provide a SQL Server database sandpit for query analysis in the ADA area. Two S2-sized Azure SQL Databases, one for OPA and one for MACA are created, and each database is limited to 250 GB of storage.

User access is controlled by SQL logins on the SQL server. The ADA VMs have connectivity to these databases.

### SQL Server Analysis Services

The SQL Server Analysis Services database server allows for a private ADA instance of SSAS to run without impacting the operational production SSAS environment.

All SSAS entities from the Production environment will be available on this server. When new versions are released into Production, the new versions will also be deployed to the ADA SSAS instance.

### Spark

As the Data Lake’s HDInsight cluster can provide Hadoop queries, there is provision for using Apache Spark in ADA using either HDInsight Spark (Microsoft’s implementation of Spark) or an IaaS deployment.

As HDInsight Spark clusters can be created on demand as configuration is stored within an Azure SQL Database, and HDInsight can access the Windows Azure Storage Blobs (WASB) of the Data Lake, an HDInsight option will be deployed.

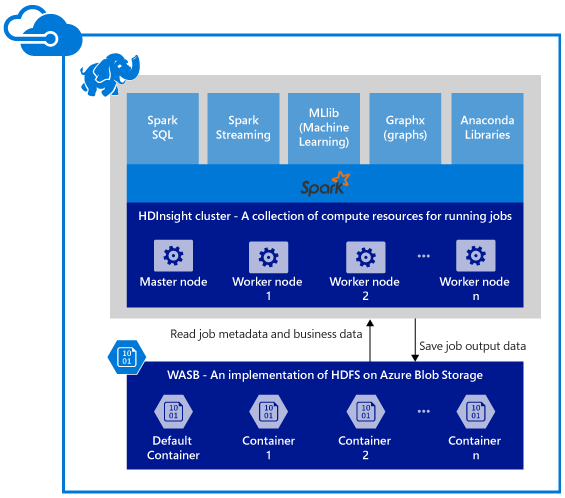


Figure Microsoft HDInsigh Spark Cluster Architecture

### Azure Machine Learning

Azure Machine Learning (Azure ML) is a cloud-based service that allows users to create machine learning experiments that load, cleanse, and process data according to machine learning algorithms. Azure ML is currently not available in Australia, but Azure ML Standard subscriptions can process data from Australia – data is not stored on the remote server, only processed.

From an architectural perspective, no additional infrastructure is required to use Azure ML. The Azure ML workspace is configured with Windows Azure Storage Blobs and Azure SQL Data Warehouse as data sources.

The choice of whether to use Spark or Azure Machine Learning for Machine Learning tasks will come down to personal choice by the data scientist, and whether the streaming of data to the US-based Azure ML servers is acceptable. As a full PaaS offering, there is no infrastructure needed to deploy Azure ML.

# Integration Architecture

The following diagram shows the Integration architecture for the DAP:

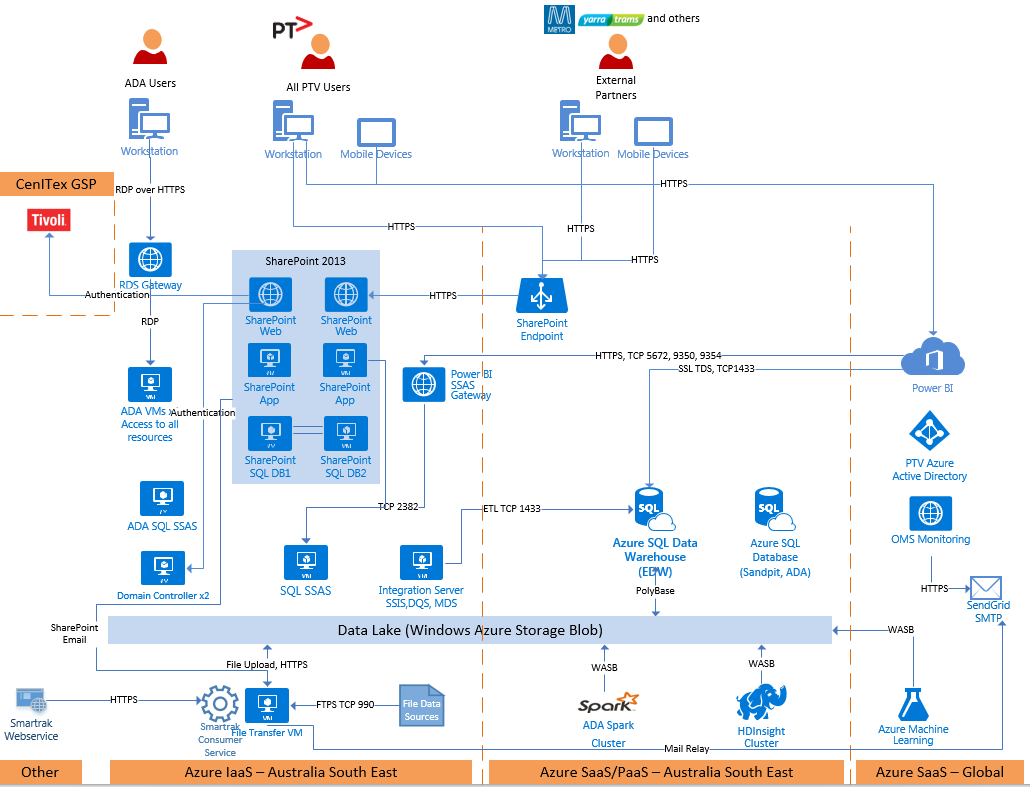


Figure Integration Architecture

The following integration items are shown above:

* As a general rule all communications is performed using secure protocols. The only use of unsecured HTTP is to initial requests to SharePoint prior to redirecting to an HTTPS connection.
* SMTP – SMTP based email is used for operational management and alerting (performance and service availability alerts), as well as for SharePoint notifications, such as workflow approval items. Emails are sent via the SMTP relay hosted on the File Transfer VM, and forwarded through to SendGrid for distribution.
* Data Platform Integration – The Microsoft SQL Server stack provides the glue between the layers of the data platform, primarily driven and scheduled via SQL Server Integration Services.
* Power BI – PowerBI has direct access to Azure cloud-based data sources, such as the EDW (Azure SQL Data Warehouse) and the ADA Sandpit databases. For other data sources, the Enterprise Gateway and Power BI SSAS Gateway are required to present an endpoint for Power BI to connect to these data sources.
* Data Lake – Integration of components accessing the Data Lake are provided by Windows Azure Storage Blob (WASB) access to the Data Lake’s storage. In the case of HDInsight (Spark or Hadoop), this is provided by Microsoft’s HDFS driver, which provides a virtual HDFS layer over the top of WASB.
* SharePoint authentication of internal users is performed by an endpoint on the CenITex GSP Tivoli Federated Identity Management (FIM) that allows connections from SharePoint. This allows single sign-on for pre-authenticated VicGov users, or challenged access for currently unauthenticated VicGov users.

# SharePoint 2013 Architecture

An on-premises SharePoint 2013 implementation in Azure has been chosen over SharePoint Online due to the SharePoint portal requiring rich SQL Server Reporting Services-based reporting. SSRS and other analytical features are not available in SharePoint online, and SharePoint Online does not allow external users to use Power View, PowerPivot or Excel Services.

The farm type in this design is a medium size on MS Azure infrastructure with High Availability. It consists of six (6) servers. In conjunction with these dedicated SharePoint 2013 server roles, the farm requires a clustered database arrangement which will provide an Active/Passive array for database services to the farm.

Based on the potential usage of the Internal SharePoint farm, a server farm consisting of two web front ends, two application servers and a SQL Cluster service with two nodes have been designed with the following topology:

* Two front-end Web servers (Load balanced using Azure Network Load Balancing) to provide scalability and high availability of the web sites. These two servers will also provide the query service and provide the ability to move roles to dedicated servers for performance reasons. Both will host the Central Administration Console.
* Two SharePoint application servers to provide scalability and high availability of the SharePoint application services. Along with SharePoint application services, these two servers will also store Search Index files on a dedicated disk drive and will be configured with SQL reporting services.
* SQL Server 2014 facility using SQL Database Cluster with two nodes.

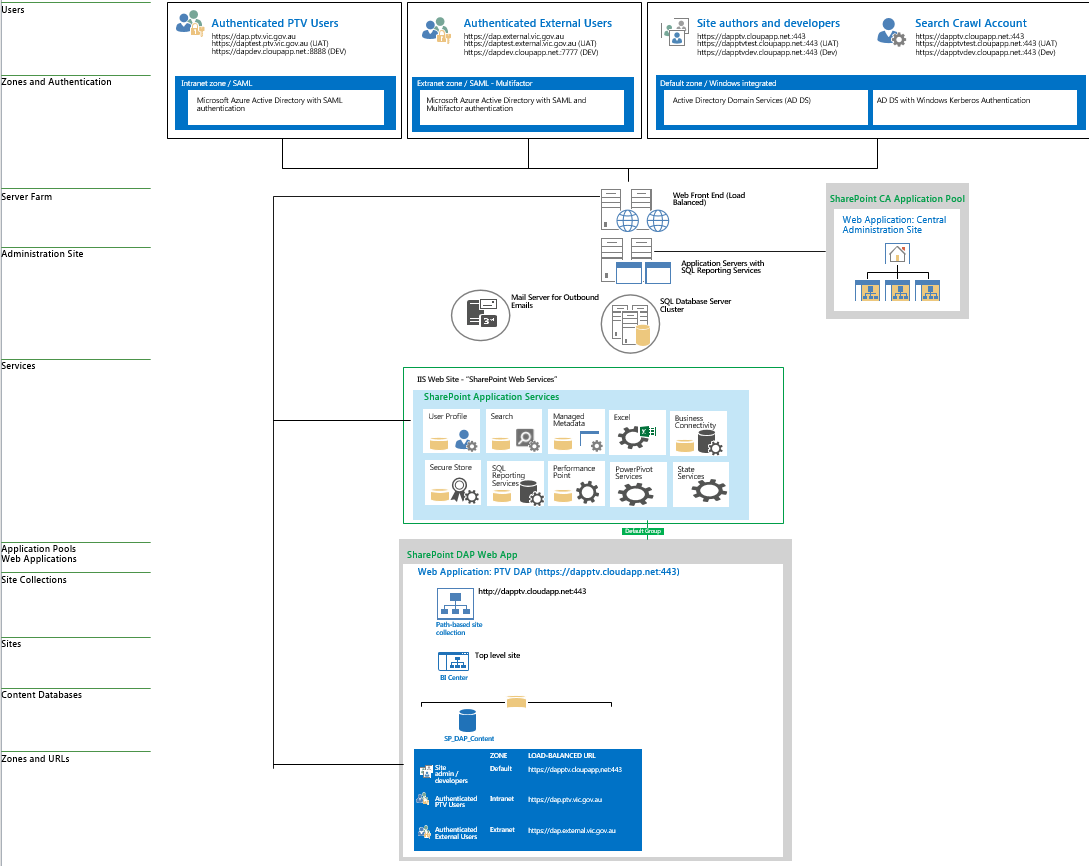
The farm has been designed to scale-out and scale-up to support future Microsoft SharePoint 2013 solutions.

This topology enables additional servers and resources to be added and roles to be reallocated as the farm grows to meet changing business needs.

It is important to note that there will need to be some careful monitoring of the farm in the early days as the users adopt the platform. The proposed topology is ideal as a starting position but may require additional resources quickly.

A significant benefit of this farm design is the ability to add a new server to assume all the services which are currently working on the two WFE role or Application servers. In this case, there is the third server offering just the WFE Role with high Affinity in the Azure Load Balancer.

All parts of the SharePoint portal will adhere to WCAG 2.0, with at least an AA conformant rating.

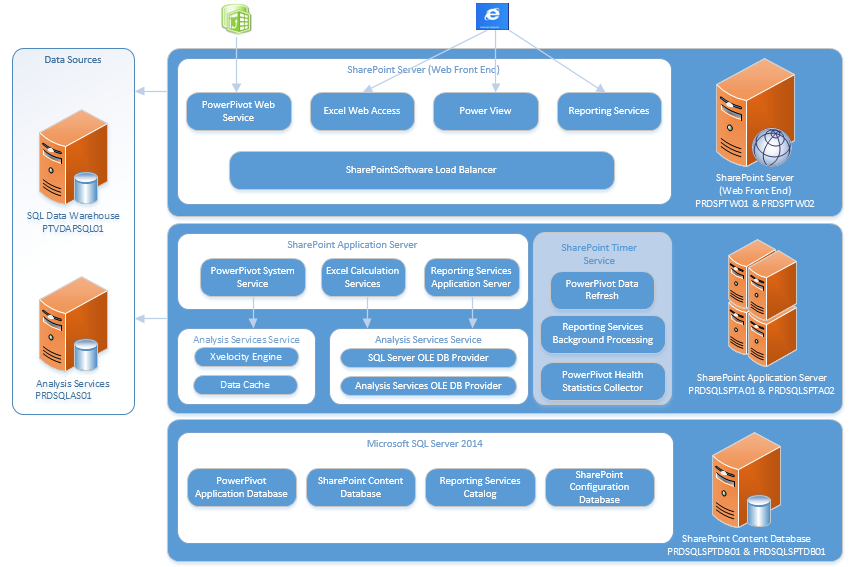


*Figure 17 – Azure SharePoint Farm*

# Presentation (Reporting) Architecture

### Reporting Presented through SharePoint 2013

SharePoint 2013 presents three main forms of reporting: SQL Server Reporting Services, Power View and Excel Services.



#### SQL Server Reporting Services

Microsoft SQL Server Reporting Services is a server-based reporting platform for creating and running reports, it can source data from both relational and multidimensional sources. Reporting Services has a deep and rich feature set that allows extensive customisation and integration with SharePoint.

**Note:** Reporting Services is geared towards professional report authors, who are usually specialist BI developers.

SSRS reports presented through SharePoint are the location where operational information about the EDW load processes are stored, and can be reviewed by support and authorised users.

#### Power View

Power View is an easy to use dash boarding tool that enables self-service business intelligence. Users can quickly and easily create (and share) reports to visualize and explore data. Power View can connect to PowerPivot workbooks (uploaded to SharePoint) or to Tabular and Multidimensional databases. PowerView has migrated from a Silverlight application that runs through SharePoint to being embedded in Excel 2013. Both variants are supported by the DAP, but the Excel 2013 version will slowly phase out the Silverlight application as the PTV users migrate off Office 2010.

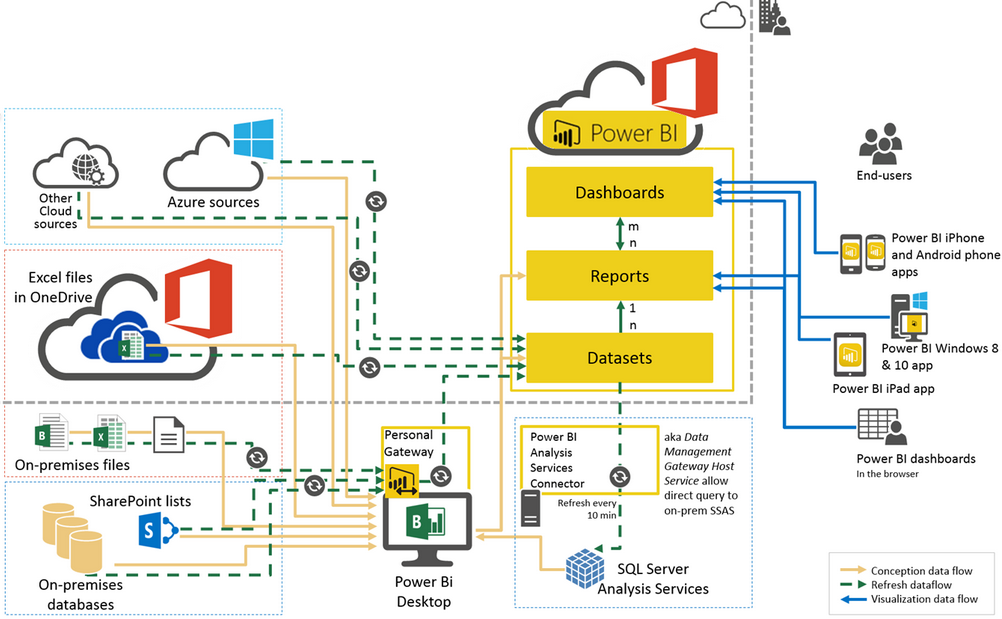
#### Excel Services

Excel Services is a service application that enables users to load, calculate, and display Microsoft Excel workbooks on Microsoft SharePoint 2013. Users can create, view, interact, and share Microsoft Excel workbooks from within a web browser. Excel Services in SharePoint 2013 also includes data model functionality to enable interaction with Power Pivot (models) workbooks from within a web browser.

### Power BI

Power BI is used in the DAP to provide an interface for PTV staff to engage in self-service BI and sharing with other members in the organisation.

Power BI’s architecture is shown in the following diagram. While this diagram is a generalised architecture diagram, PTV will be using all components other than OneDrive. “On-Premises” can be read as either “PTV Premises” or “Within the DAP”, although only the latter (DAP) will be used for Proof of Value. Installation of the Personal Gateway is outside the scope of the DAP, but some Power BI users may find it useful.



Power BI is a cloud-based reporting solution, currently hosted in the US. Microsoft has advised that they will (in January 2016) release the ability to create an Australian-based Power BI tenancy, and PTV will be migrated to this during February 2016.

There are three main ways of using Power BI, and all three will be used by the DAP:

* Upload data sources to Power BI (typically Excel spreadsheets). This is available with Power BI Free, but data sources can be refreshed no more often than once every 24 hours. Datasets uploaded to Power BI Free are stored on the Australian Power BI servers.
* Connect to a non-Power BI data source – this includes the SQL Data Warehouse, and an on-premises Analysis Services server – in the case of the DAP, SSAS is run as an “on-premises” deployment in the cloud. These features require Power BI Pro.
* Power BI Desktop can be used locally without needing to upload any data.

Power BI hosts the report/visualisation definitions on the Power BI servers, as well as the definition of data sources.

The second option listed above (requiring the Pro subscription) reads the report/visualisation definitions, and connects to the data sources. It (Power BI in the US) then queries the data source, packages the data into the visualisation and returns the report. No data for these live-refresh reports are stored or cached.

The DAP PoV budget allows for 175 Power BI Pro users during Year 1, progressing to 235 users in Year 3. The primary reason for Power BI in the DAP is to provide self-service BI capabilities rather than static reporting, and those reports that will go out to large numbers of users are to be created as SQL Server Reporting Services reports and presented through SharePoint. Note that a future upgrade (out of the scope of the Proof of Value) to SQL Server 2016 will allow for SSRS reports to be presented through Power BI, further blurring the lines between SharePoint and Power BI.

Power BI is currently organisation-specific, in that reports cannot be shared between users in different organisations. This limitation is currently the most requested feature, and Microsoft are working on a solution. Until that time, external users will be restricted to curated reports provided through SharePoint, or granted access to the ADA environment to perform analysis.

Power BI, including the Power BI portal, is a Microsoft COTS product and adherence to accessibility guidelines such as WCAG are under the control of Microsoft.

#### Reporting Tool Feature Comparison

The following table shows the intended audience for creating reports with each reporting tool:

|  |  |  |  |
| --- | --- | --- | --- |
| Audience Type | SSRS | Power BI | Excel Services |
| BI/Reporting Developers | ✔ | ✔ |  |
| Power Users |  | ✔ | ✔ |
| BI Self-Service Users |  | ✔ | ✔ |
| Data Exploration Users |  | ✔ |  |

The following table lists benefits for PTV for each type of reporting technology:

|  |  |  |  |
| --- | --- | --- | --- |
| Benefit | SSRS | Power BI | Excel Services |
| Provide authoritative sources of information which will support PTV's reporting functions and inform business decision making processes | ✔ | ✔ |  |
| Reduce the manual effort needed to produce reports |  | ✔ | ✔ |
| Self-Service BI |  | ✔ | ✔ |
| Increase the range of business metrics that PTV can report on | ✔ | ✔ |  |
| Improve visibility into the reporting of PTV's business performance | ✔ | ✔ |  |
| Improve the accuracy of reporting with reduced risk of error | ✔ | ✔ |  |
| Increase the frequency and timeliness of reporting activity | ✔ | ✔ |  |
| Deliver enhanced visualisation of reporting and statistical dashboard capabilities | ✔ | ✔ | ✔ |
| Support a significant (exponential) growth in the quantity of disaggregated reporting |  | ✔ |  |
| What-If Scenarios |  | ✔ |  |
| Key performance Indicators |  | ✔ | ✔ |
| Excel based Reporting |  |  | ✔ |
| Maps Reporting requirements | ✔ | ✔ | ✔ |
| View Reports on Mobile Devices | ✔ | ✔ |  |

# Technology Architecture

## Environments

The following environments will be deployed.

| Environment | Purpose | Platform | Created | Lifecycle | Description |
| --- | --- | --- | --- | --- | --- |
| Dev | Development | Azure | Nov-15 | Permanent | Reduced capacity  Minimal redundancy |
| Test | Testing | Azure | Dec-15 | Permanent | Reduced capacity  Full redundancy to mimic production |
| Prod | Production | Azure | Jan-16 | Permanent | Full capacity  Full redundancy |
| DR | Disaster recovery | Azure | Jan-16 | Activated when needed | Full capacity  Full redundancy |

**18 Environments**

In order to reduce cloud computing costs, the development and test environments will have restricted hours of operation once the solution is fully implemented in year 2. After this time, the development and test environment is expected to be required for 5 days every month.

| Environment | Operation | Year1 | Year2 | Year3 |
| --- | --- | --- | --- | --- |
| Dev | Reduced | Daily  15 hours/day | 5 days/month  15 hours/day | 5 days/month  15 hours/day |
| Test | Reduced | Daily  15 hours/day | 5 days/month  15 hours/day | 5 days/month  15 hours/day |
| Prod | Production | Daily  24 hours/day | Daily  24 hours/day | Daily  24 hours/day |

**19 Environment operation**

Within the DAP solution, three main services are to be delivered.

* ADA – Analytics Discovery Area
* DP – Data Processing
* SP – SharePoint

Although there are three discrete environments, some communication will still initially be possible between the environments, including shared infrastructure, such as Active Directory, and servers used to manage the environment (jump hosts). As a large proportion of the environment is cloud-based Software as a Service components, it is difficult to prevent some servers from gaining access, and would require a more advanced firewall to be deployed that can filter based on hostname, not IP address. For example, to prevent traffic from the Dev ETL server (DEVSQLETL01) connecting to the Production EDW, a rule would need to be created to block traffic to PTVDAPSQL01.database.windows.net, regardless of the IP address. This advanced firewall may be deployed as part of the future enhanced security updates to the DAP.

Until this time, environments are protected from accidental access by different user credentials in each environment.

## Major Software Versions and Licensing

The following major software versions are used by the DAP:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Software Name | Version | Vendor | License Model | Licensing Scheme | # Licenses - Prod | # Licenses - Test | # Licenses - Development |
| Windows Server 2012 | 2012 R2 Datacenter | Microsoft | Azure VM Consumption | N/A – the price of licensing is built into the hourly cost of running the VM. | - | - | - |
| SharePoint 2013 | 2013 Enterprise with SP1 | Microsoft | Server License + User CALs | 4xServer Licenses in Production. User CALs provided by Enterprise Agreement for PTV staff, external users do not require a license. | 4xServer | MSDN | MSDN |
| SQL Server (IaaS) | 2014 SP1, CU3 | Microsoft | Azure VM Consumption | N/A – VMs running a SQL Server IaaS implementation are built on a VM image that has a per-hour pricing uplift to include costs for the SQL licenses. No additional licenses are included. | - | MSDN | MSDN |
| Azure SQL Database (DBaaS) | V12 | Microsoft | Azure Consumption | N/A | - | - | - |
| Azure HDInsight | 3.2 (Hadoop 2.6.0) | Microsoft | Azure VM Consumption | N/A | - | - | - |
| Power BI Pro | N/A | Microsoft | Per User |  | Year 1: 175 Year 3: 235 | - | - |
| SQL Server Management Studio/Data Tools | 2014 | Microsoft |  | As long as users are connecting to properly licensed production SQL Servers, there is no license requirement for SSMS/SSDT | - | - | - |
| R Studio | Open Source | R Studio, Inc | Open Source (AGPL v3) or Commercial | N/A – Open Source version to be used. | - | - | - |
| Microsoft Excel | 2013 | Microsoft | Virtual Desktop | N/A – Excel on a remote host is considered licensed as long as the desktop/laptop initiating the remote connection is licensed to run Excel. | - | - | - |
| Azure Active Directory (DAP AAD) | N/A | Microsoft | Free Tier | N/A | - | - | - |
| Multi-Factor Authentication | N/A | Microsoft | Per User or Per 10/Authentications | Multi-factor Authentication is applied at the Azure AD level, for the DAP AAD (External Users only). Pricing is US$1.78 for each user per month (known amount per month), or the same for 10 authentications for any user (variable cost per month). |  | - | - |
| Certificates | N/A | GeoTrust | Per License | 3 certificates required – FTPS, HTTPS (SharePoint) and RDS. | Extended Validation Certificate | Domain Validated Certificate | No Certificate |
| Remote Desktop Services | N/A | Microsoft | Per Concurrent User (not named user) | 40 licenses required over the ADA and Development/Management environment. | 40 | | |
| Operations Management Suite (OMS) | Azure | Microsoft | Consumption – per GB | Cost depends on size of stored logs – first 500 MB is free, then $3.50 per GB. | - | - | - |
| SMTP Service | Azure | SendGrid | Tiered – Free Tier | The free tier allows 25,000 emails to be sent per month. | - | - | - |

### Licensing for Non-Production

As described in the table above, non-production is licensed via Microsoft Developer Network (MSDN) licenses. It is possible to create a separate Azure subscription marked as MSDN that gets a discount on running costs. Microsoft recommend using the smallest number of subscriptions possible, expanding to multiple subscriptions only for special billing requirements that can’t be met with standard reports or tagging items, exceeding the limits for a single subscription, requiring administration separation that cannot be achieved via the built in RBAC controls, or to achieve a cheaper cost for non-production via an MSDN subscription.

The largest cost that MSDN avoids is the cost of SQL Server and SharePoint Server licenses, and MSDN versions are installed in the Development and Test regions. The main impact is on paying for Windows Server licenses, which is approximately 60% more expensive that the MSDN option. In the development environment, as servers run 15 hours a day, this results in approximately $1064 in monthly costs under a full subscription, compared to $902 for an MSDN subscription, resulting in approximately $160 per month. As the Test environment consists of twice the number of virtual machines, this saving can be doubled, resulting in $480 per month that could be saved if the Development and Test environments had an MSDN subscription. Out of the total monthly Azure spend of $27,000 per month, this would result in a 1.7% saving.

The decision to use a single subscription was to keep management overheads down, as an Azure virtual network cannot span multiple subscriptions, and would require either additional copies of management servers (Windows Update, Active Directory), or routing between the two environments. The additional cost of these components would offset any potential saving by moving to a separate MSDN subscription for Development and Test.

## Network View

### Connectivity

The Azure platform provides multiple connectivity options for clients. As much as possible, the connectivity options employed for DAP will maintain a high separation of network connectivity between the DAP hosted solution and the PTV private networks. Options such as site-to-site VPN over the Internet or ExpressRoute as an extension of the PTV private network will be employed as they become available but will be firewalled from PTV private networks.

The following connectivity options will be provided for the DAP solution:

* Internet to DAP services
* Site-to-site VPN
* ExpressRoute to DAP services

It was initially proposed that all PTV connectivity would be provided using the ExpressRoute transport. This will ensure that the quality of service for the connectivity to the DAP services can be controlled to a greater degree than when accessed over the Internet.

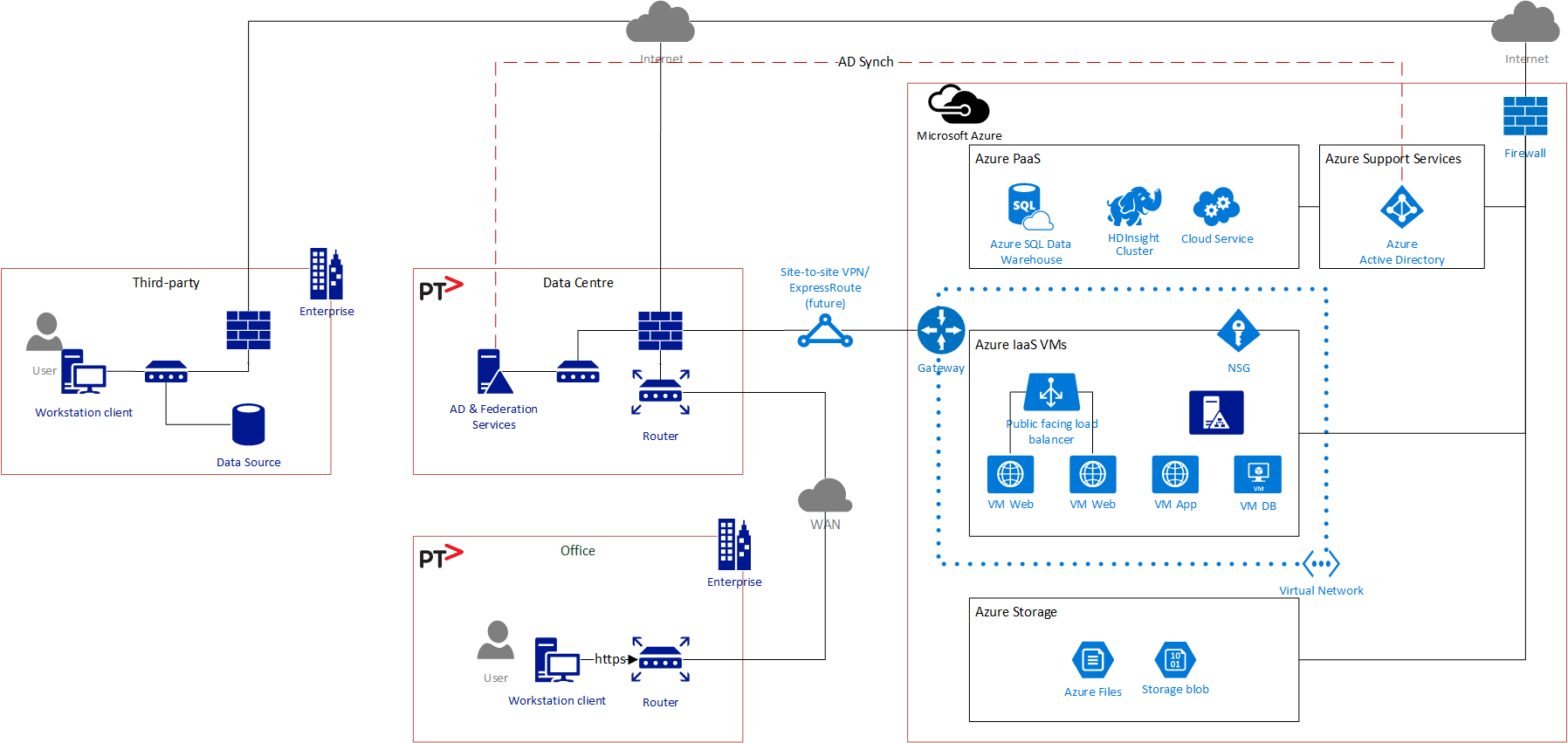
CenITex is currently in the process of planning for the implementation of ExpressRoute connectivity for Victorian government agencies. This is expected to be available next year. Due to the security complexity associated with new network connections to the PTV network, the implementation of ExpressRoute for DAP will be postponed until the CenITex services can be leveraged. Initial connectivity to DAP will be provided over the internet with a site-to-site VPN being deployed over the course of the build phases. ExpressRoute will be deployed to replace the site-to-site VPN when it becomes available.

The network connectivity use cases for the DAP solution will be as follows.

| Role | Transport (day 1) | Transport (future) | Service |
| --- | --- | --- | --- |
| PTV users – internal | Internet | Site-to-site VPN/ ExpressRoute | Web services |
| PTV ADA users - internal | Internet | Site-to-site VPN/ ExpressRoute | Web services  Remote desktop services |
| PTV operators | Internet | Internet | Web services data transfers |
| Third-party data sources | Internet | Internet | Web services or FTPS data transfers |

**20 Network connectivity use cases**

An overview of the network connectivity for the DAP solution is as follows.



**21 Network connectivity overview**

#### ExpressRoute

This will be provided as a future connection option once CenITex provide this as a shared service to the Victorian Government.

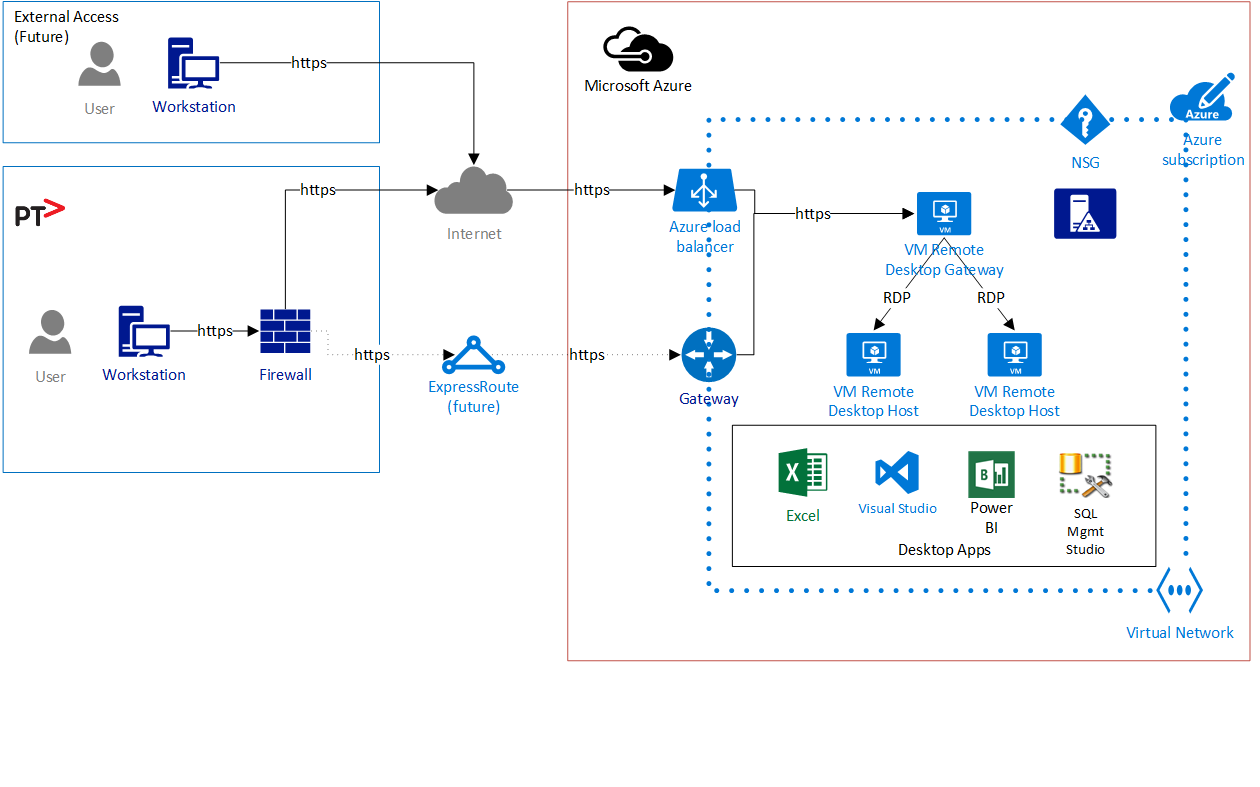
ExpressRoute utilises fibre connections provided by telecommunications providers to provide point to point connections between Azure subscribers’ local networks and the Azure data centres. These connections are configured by Microsoft to provide direct internal connections to Office 365, Azure PaaS services and the Azure subscriber’s IaaS networks. ExpressRoute subscribers can choose which of these connections (peerings) to connect but in the case of a shared ExpressRoute connection all IaaS subscriptions using the same circuit are able to communicate with each other.

It is anticipated that CenITex will require specific routes and services to be configured to prevent this happening. However, these requirements will not be known until CenITex have completed the ExpressRoute deployment.

This document will be updated to reflect these requirements once they are known.

#### Remote Desktop Services

Secure desktop access to the DAP environment will be provided using Windows Remote Desktop Services Gateway and Windows Remote Desktop Session Hosts. This will allow PTV users to initiate secure remote desktop sessions from the PTV internal network using https. Initially this connectivity will be over the Internet. In the future, this will be delivered over ExpressRoute. Any future access to the Remote Desktop environment for external users will be provided over the internet using the same https connection.



**22 Remote desktop connectivity - PTV**

Similarly, CGI will use secure remote desktop services to administrator and support the DAP environment. Internally within CGI, this access is delivered using an SSL gateway hosted in our isolated ServiceHUB support environment. This ensures a high degree of separation between CGI corporate networks and the client networks that CGI support.



**23 Remote desktop connectivity – CGI support**

### Data Transfers

There will be two use cases for network connectivity for data transfers:

* Programmed transfers to Azure Storage Accounts
* Interactive transfers using FTPS

For trusted data transfers from PTV to the Data Lake, direct transfers to Azure Storage accounts will be utilised. Initially the transport will be https over the Internet with the option to move to ExpressRoute in the future.



**24 Data transfer connectivity – storage accounts**

For interactive data transfers from third parties, a secure FTP server will be hosted as an Azure VM. This will use Azure File shared storage to enable the data to be transferred within the DAP environment to the Data Lake. The FTPS service will be published using a reserved IP and accessible via a known URL.



**25 Data transfer connectivity – FTPS**

### Network Topology

The DAP solution within the Azure Infrastructure Services hosted environment will be deployed using a single Azure virtual network (VNet). This will be divided into a classic three tier network topology for web, application and database. The network will be further separated for dev, test and production environments. Each network will have a network access policy applied using Azure network security groups. This will allow network traffic to be restricted between each network zone.

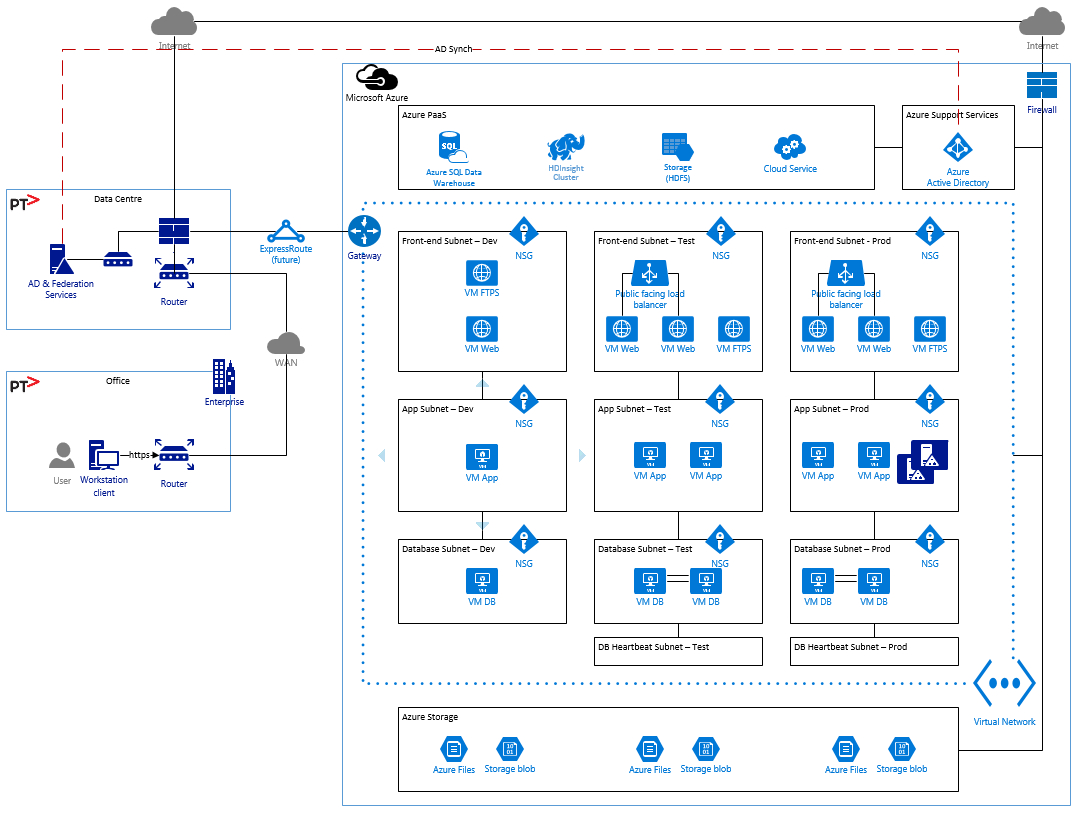
Azure PaaS applications and storage services will be separately consumed from the Azure fabric.

The networks to be created are as follows.

| Fabric/Virtual Network | Subnet | Purpose |
| --- | --- | --- |
| Azure Support Services | Azure Support Services | Azure Active Directory |
| Azure PaaS | Azure PaaS | Shared service instances for SQL data warehouse, HD Insight, Cloud Services |
| DAP VNet (Single VNet with separate subnets for Dev/Test/Prod) | Web Dev | DAP web services |
| App Dev | DAP application services |
| Database Dev | DAP database services |
| DB Heartbeat Dev | DAP database heartbeat (allocated but not used) |
| Web Test | DAP web services |
| App Test | DAP application services |
| Database Test | DAP database services |
| DB Heartbeat Test | DAP database heartbeat |
| Web Prod | DAP web services |
| App Prod | DAP application services |
| Database Prod | DAP database services |
| DB Heartbeat Prod | DAP database heartbeat |
| Shared | Management Services, Active Directory |
|  | ADA | DAP ADA Workstations |
|  | UnTrust | Firewall UnTrusted subnet |
|  | Trust | Firewall Trusted subnet |
|  | NAT | NAT Subnet to provide public IP for firewall UnTrust |
| Azure storage | Azure storage services | Azure storage services |

**26 DAP networks**

An overview of the DAP networks is as follows.



**27 DAP network overview**

### Network Addressing

All network addressing for both private and public addressing will provided within the Azure environment, in isolation of any existing PTV addressing. The private addressing will be configured to use the ranges provides by CenITex.

| Address Type | Address Space | Allocation | Description |
| --- | --- | --- | --- |
| Public | 40.x.x.x | Dynamic | Azure registered addressing |
| Private | 10.209.0.0/24 | Dynamic with reservations as required | Private addressing |

**28 Network addressing**

Further details on the subnets that will be used within the private address range provided by CenITex can be found in Appendix 1 at the rear of this document.

### Network Services

All network services within the DAP hosted environment will be provided by Azure services. This will include the following.

| Service | Resource |
| --- | --- |
| DNS | Internal domain - hosted Active Directory  External domains – Azure DNS |
| Network access control | Azure Firewall  Azure Network Security Group |
| Network Load Balancer (NLB) | Azure Cloud Services and Azure Network Load Balancer |
| Time synchronisation | Azure platform time synchronisation |

**29 Network services**

### Network Security Groups

Network Security Group (NSG) allows the creation of rules (ACLs) at the desired level of granularity: network interfaces, individual VMs, or virtual subnets. Access can be controlled by permitting or denying communication between the workloads within a virtual network, from systems on customer’s networks via cross premises connectivity, or direct Internet communication.

For DAP, NSG policies will be applied to the virtual subnets to restrict network traffic in both north-south (Internet, web, application, database) and east-west (dev, test, production) directions. The NSG policies will be based upon the following logical connectivity.

| Source | Destination | Service |
| --- | --- | --- |
| Internet | Web Dev, Web Test, Web Prod | https, ftps |
| Web Dev | App Dev | SharePoint application |
| App Dev | DB Dev, SQL PaaS, HD Insight | SQL Server |
| Web Test | App Test | SharePoint application |
| App Test | DB Test, SQL PaaS, HD Insight | SQL Server |
| Web Prod | App Prod | SharePoint application |
| App Prod | DB Prod, SQL PaaS, HD Insight | SQL Server |
| Web Dev, App Dev, DB Dev | Shared | Active Directory, Monitoring |
| Web Test, App Test, DB Test | Shared | Active Directory, Monitoring |
| Web Prod, App Prod, DB Prod | Shared | Active Directory, Monitoring |
| Web Prod | Web Dev, App Dev, DB Dev | Remote Desktop Services |
| Web Prod | Web Test, App Test, DB Test | Remote Desktop Services |
| Web Prod | Web Prod, App Prod, DB Prod | Remote Desktop Services |

**30 NSG logical policy**

## Infrastructure Architecture

### Overview

The DAP solution will be hosted on the Azure platform in Australia. The primary location will be Azure Australia South-East and the secondary (DR) location will be Azure Australia East.

This will be provided using a new Azure subscription specific to the DAP project, created under the PTV organisation account. This will allow CGI to be the primary administrator of the Azure subscription and be responsible for the configuration and service delivery. Future PTV projects can then operate as an additional Azure subscription, independently from the DAP environment, if required.

The DAP solution will consume three main Azure service types:

* Azure Platform-as-a-Service (PaaS) applications
* Azure Infrastructure-as-a-Service (IaaS) – virtual servers (VMs)
* Azure Storage Services

All Azure components of the DAP provide at least 99.5% availability, and Availability Sets are used to reduce single points of failure.

Estimated consumption (in dollars) of each Azure component is available in the “DAP Estimated Consumption” attachment at the end of this document.

### Resource Groups

Azure Resource Manager provides the capability to organise the Azure resources into Resource Groups. The benefits of these Resource Groups to the DAP are that:

* Resources for the DAP solution are deployed, managed, and monitored as a group, rather than handling these resources individually.
* Components of the DAP are deployed throughout the development lifecycle and in DR activities in a consistent state.
* Declarative templates are used to define the deployment to DR.
* Access control to services in Resource Groups can be controlled at a group level, and integrated into the Azure Portal management platform. At the current time, no access to the Azure Portal is available to anyone other than DAP system administrators (CGI and PTV), but some ADA users may be granted access to control specific components, based on the Resource Group.
* Tags are be applied to some resources to logically organise the resources in the subscription for grouping and billing purposes. At the current time, the only tags to be applied have been to mark specific ADA virtual machine components in addition to their resource group.

For DAP, separate resources groups will be used for each environment as follows.

| Resource Group | Environment | DAP Services |
| --- | --- | --- |
| DAP-Dev | Development | Data Processing |
| DAP-Dev-ADA | Development | ADA |
| DAP-Dev-SP | Development | SharePoint |
| DAP-Test | Test | Data Processing |
| DAP-Test-ADA | Test | ADA |
| DAP-Test-SP | Test | SharePoint |
| DAP-Prod | Production | Data Processing |
| DAP-Prod-ADA | Production | ADA |
| DAP-Prod-SP | Production | SharePoint |
| DAP-Shrd | Shared | DAP Active Directory  Management |
| DAP-Security | Shared | Environment Security Resources |

**31 DAP resource groups**

### Azure PaaS

The Azure PaaS applications provide preconfigured application and database services that can be deployed rapidly, decreasing the implementation time and effort. DAP will consume the following application services:

* SQL database services
* HD Insight
* Cloud services – network load balancer (NLB)

### Azure IaaS

Where more specific configurations or third-party applications are required, Azure Infrastructure Service servers will be consumed. For DAP, the uses will include:

* Active Directory servers
* SharePoint web, application and database servers
* Custom Analytics
* Remote Desktop Services

All Azure VMs deployed for DAP will utilise Windows Server 2012 R2 operating system.

### Azure Storage

For both the Azure application-as-a-service and Azure VM services storage will be consumed from the Azure storage services. This will include:

* Storage blob – VM virtual hard disk (VHD), database data, data warehouse data lake
* Azure files – file shares

Azure offers multiple levels of storage redundancy:

* LRS – locally redundant storage
* ZRS – Zone-redundant storage
* GRS – Geo-redundant storage
* RA-GRS – Read-access geo-redundant storage

The redundancy and replication strategy for these storage levels are as follows.

| Replication Strategy | LRS | ZRS | GRS | RA-GRS |
| --- | --- | --- | --- | --- |
| Data is replicated across multiple facilities | No | Yes | Yes | Yes |
| Data can be read from the secondary location as well as from the primary location | No | No | No | Yes |
| Number of copies of data maintained on separate nodes | 3 | 3 | 6 | 6 |

**32 Azure storage replication**

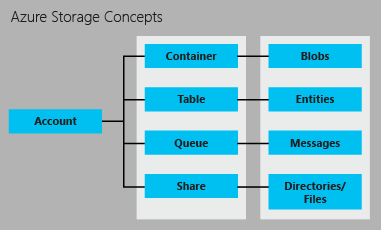
For DAP, two levels will be used:

* LRS – for development and test with local redundancy only
* GRS – for production with geo-redundancy for disaster recovery

Where GRS is used the primary copy will be located in the Australia South East DC and replicated to the Australia East DC.

### Azure Storage Accounts

Azure storage is defined within storage accounts. Each storage accounts may consist of a collection of all types of Azure storage.



**33 Azure storage concepts**

In general, a storage account will be created for each resource group. Where the storage requires specific access requirements, such as for the data transfer storage, a separate storage account will be created. This will allow for the access to be securely managed.

| Resource Group | Storage Account | Storage Type | DAP Services |
| --- | --- | --- | --- |
| DAP-Prod | ptvdap01 | Blobs | DP VMs |
| ptvdap01datatrans | Files | FTPS  HDFS |
| DAP-Prod-ADA | ptvdap01ada | Blobs | ADA VMs |
| DAP-Prod-SP | ptvdap01sp | Blobs | SharePoint VMs |
| DAP-Test | ptvdap02 | Blobs | DP VMs |
| ptvdap02datatrans | Files | FTPS  HDFS |
| DAP-Test-ADA | ptvdap02ada | Blobs | ADA VMs |
| DAP-Test-SP | ptvdap02sp | Blobs | SharePoint VMs |
| DAP-Dev | ptvdap03 | Blobs | DP VMs |
| ptvdap03datatrans | Files | FTPS  HDFS |
| DAP-Dev-ADA | ptvdap03ada | Blobs | ADA VMs |
| DAP-Dev-SP | ptvdap03sp | Blobs | SharePoint VMs |
| DAP-Shrd | ptvdapshrd | Blobs | Mgmt VMs  AD VMs |
| DAP-Security | ptvdap01sec | Blobs | Firewall VMs |

**34 DAP storage accounts**

### Availability Sets

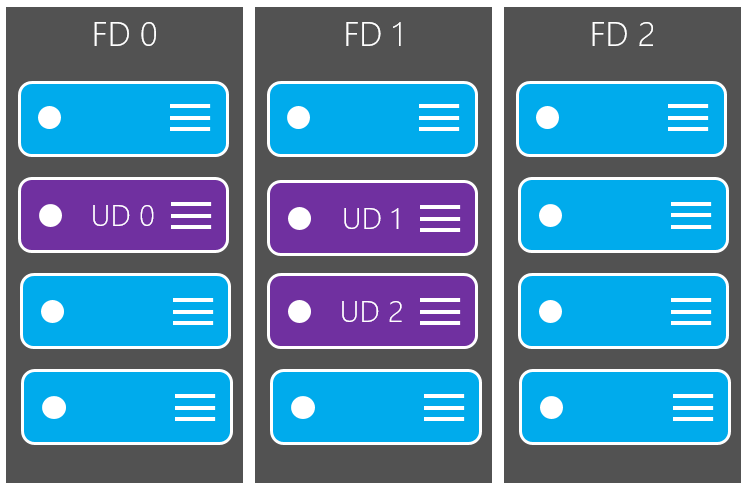
Azure availability sets provide increased redundancy for groups of virtual servers. This protection is provided for two types of events:

* Planned maintenance – periodic updates made by Microsoft to the underlying Azure platform to improve overall reliability, performance, and security of the platform infrastructure that your virtual machines run on where a reboot may be required
* Unplanned maintenance when the hardware or physical infrastructure underlying your virtual machine has faulted in some way.

Each VM in an availability set is assigned an Update Domain (UD) for planned maintenance and a Fault Domain (FD) for unplanned maintenance.

For update domains, virtual machines in the same domain will be restarted together during planned maintenance. Azure never restarts more than one update domain at a time. VMs in an availability set are spread across up to 5 update domains.

For fault domain, virtual machines in the same domain share a common power source and physical network switch.



**35 Availability sets**

For DAP, the following availability sets will be used.

| Environment | Availability Set | Purpose |
| --- | --- | --- |
| *Dev* | *None* | *None* |
| Test | Test-SP-Web | SharePoint web |
| Test-SP-App | SharePoint application |
| Test-SP-DB | SharePoint database cluster |
| Prod | Prod-SP-Web | SharePoint web |
| Prod-SP-App | SharePoint application |
| Prod-SP-DB | SharePoint database cluster |
| Management | Prod-AD | Active Directory |

**36 DAP availability sets**

### Active Directory – DAP

A new Active Directory domain will be hosted in Azure to provide directory services for the Azure VMs in the DAP environment. A single domain will be used for all Azure VMs across all environments. The domain will house the following objects.

* Computer accounts for Azure VMs
* User accounts used by CGI to manage the VMs
* Service accounts for applications installed on the Azure VMs

The features of the new domain will be as follows.

| Feature | Value | Configuration |
| --- | --- | --- |
| Forest | - | Single forest |
| Domain | dap.int | Single domain |
| Sites | Azure-AU | Single site – Azure South East  DR failover to Azure East |
| Subnets |  | A single subnet will be created for the assigned Azure virtual network. |
| Functional Level – forest | Windows Server 2012 R2 |  |
| Functional Level – domain | Windows Server 2012 R2 |  |
| Trusts | None |  |
| Domain controllers | <dc1>.dap.int  <dc2>.dap.int | Two domain controllers |
| Controller placement | Azure South-East | Azure availability set |
| Replication | - | Default topology |
| Global catalog | <dc1>.dap.int  <dc2>.dap.int | Both controllers will be configured as global catalog |
| Read-only controller | None |  |
| Password replication | Default |  |
| FSMO – Schema Master | Primary - <dc1> |  |
| FSMO – Domain Naming Master | Primary - <dc1> |  |
| FSMO – Infrastructure Master | Primary - <dc1> |  |
| FSMO – RID Master | Primary - <dc1> |  |
| FSMO – PDC Emulator | Primary - <dc1> |  |
| DNS | dap.int | Hosted local domain  Forwarder for external domains |

**37 DAP active directory settings**

#### Remote Desktop

Remote Desktop Services requires logons to authenticated using traditional Windows authentication through the use of either local accounts on the RDS host or accounts in a domain trusted by the RDS host. This limits the mechanisms available for provisioning accounts to users. Remote Desktop sessions will be used in two cases, CGI staff connecting to the DAP environment for administration or development purposes or PTV staff connecting to the ADA environment.

* CGI Staff – CGI staff will use accounts provisioned in the dap.int Active Directory Domain
* ADA Users – ADA users will have accounts provisioned in the dap.int Active Directory Domain until an alternative solution con be confirmed. These accounts will be matched to their existing PTV accounts with the user responsible for keeping passwords in sync.

CGI will work with CenITex to evaluate alternatives to allow single sign on for PTV users to the Remote Desktop environment. Potential solutions include

* Allowing the dap.int domain to trust accounts from the PTV domain.
* Configuring the dap.int domain to allow the CenITex IAM solution to write synchronised accounts to the domain

### Active Directory - PTV

The Whole of Vic Government domain is selectively (@ptv.vic.gov.au accounts only) synchronised to Azure Active Directory (PTV’s Office 365 directory). This allows trusted web applications to be configured to use the Azure AD to authenticate PTV users. With PTVs existing configuration (ADConnect syncing usernames and hashed passwords) this will only provide same sign on as no federation services have been implemented. Federated authentication will be performed against Tivoli Federated Identity Manager from SharePoint, allowing PTV (or other Vic Government users) to authenticate with Single Sign On.

#### SharePoint

For external users, SharePoint will be configured to use the DAP Azure Active Directory as a trusted identity provider. SharePoint will then use SAML and Web Services Federation to validate the identity of each user access request against the Azure Active Directory.

For internal users, SharePoint will be configured to use CenITex’s Tivoli FIM as a trusted identity provider. This will provide seamless single sign-on for PTV users when they access the SharePoint portal from their PTV domain workstations.



**38 Azure Active Directory authentication for SharePoint**

### Azure Resources

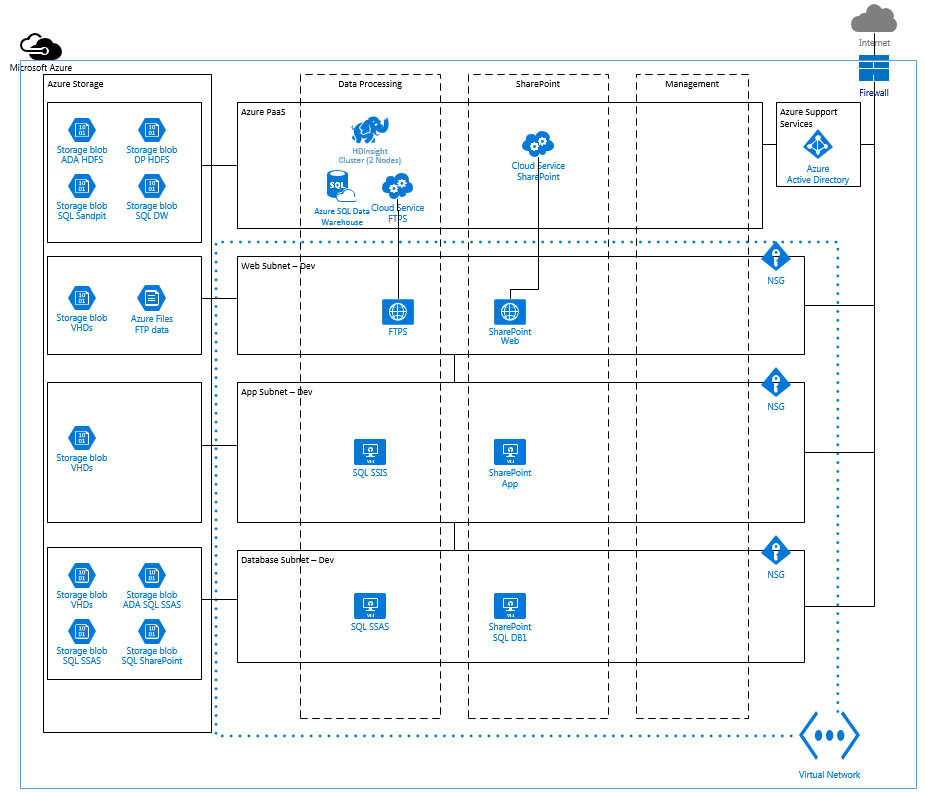
The assigned resources for each DAP environment are as follows.

#### Development

The list of assigned resources for development is as follows.

| Azure Service | DAP Service | Resource | Description |
| --- | --- | --- | --- |
| Azure Support Services | Authentication | Azure Active Directory | PTV user authentication |
| Azure PaaS |  |  |  |
|  |  |
| Data Processing | HD Insight cluster – 2 nodes | HD Insight |
| Azure SQL | Data warehouse database |
| Cloud service | FTPS data transfer NLB |
| SharePoint | Cloud service | SharePoint web NLB |
| Management | *None* | *Shared from production* |
| Azure VMs | ADA | A3 VM x 2 | Custom Analytics/RDS host |
| A3 VM | SQL server SSAS |
| Data Processing | A2 VM | FTPS server |
| A2 VM | SQL server SSIS |
| A3 VM | SQL server SSAS |
| SharePoint | A2 VM | SharePoint web |
| A2 VM | SharePoint app |
| A3 VM | SharePoint database (SQL server) |
| Management | *None* | *Shared from production* |
| Azure storage – LRS |  | | |
| Data Processing | Storage blob | SQL data warehouse |
| Storage blob | HD Insight HDFS |
| Storage blob | VM VHDs |
| Azure Files | Data transfer |
| Storage blob | SQL SSIS |
| Storage blob | SQL SSAS |
| SharePoint | Storage blob | VM VHDs |
| Storage blob | SQL database SharePoint |

**39 DAP resources – development**



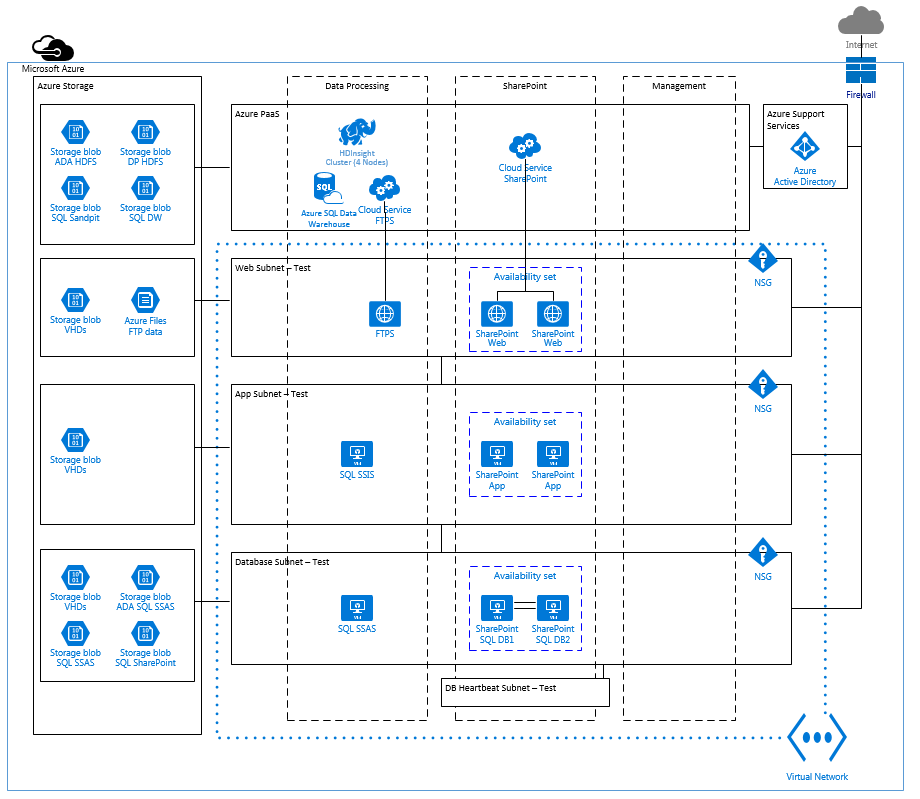
**40 DAP resource overview – development**

#### Test

The list of assigned resources for test is as follows.

| Azure Service | DAP Service | Resource | Description |
| --- | --- | --- | --- |
| Azure Support Services | Authentication | Azure Active Directory | PTV user authentication |
| Azure PaaS |  |  |  |
|  |  |
| Data Processing | HD Insight cluster – 4 nodes | HD Insight |
| Azure SQL | Data warehouse database |
| Cloud service | FTPS data transfer NLB |
| SharePoint | Cloud service | SharePoint web NLB |
| Management | *None* | *Shared from production* |
| Azure VMs | ADA | A3 VM x 2 | Custom Analytics/RDS host |
| D3 VM x 2 | Custom Analytics/RDS host |
| A3 VM | SQL server SSAS |
| Data Processing | A2 VM | FTPS server |
| A2 VM | SQL server SSIS |
| A3 VM | SQL server SSAS |
| SharePoint | A3 VM x 2 | SharePoint web |
| A4 VM x 2 | SharePoint app |
| D12 VM x 2 | SharePoint database (SQL server cluster) |
| Management | *None* | *Shared from production* |
| Azure storage – LRS |  | | |
| Data Processing | Storage blob | SQL data warehouse |
| Storage blob | HD Insight HDFS |
| Storage blob | VM VHDs |
| Azure Files | FTP data |
| Storage blob | SQL SSIS |
| Storage blob | SQL SSAS |
| SharePoint | Storage blob | VM VHDs |
| Storage blob | SQL database SharePoint |

**41 DAP resources – test**



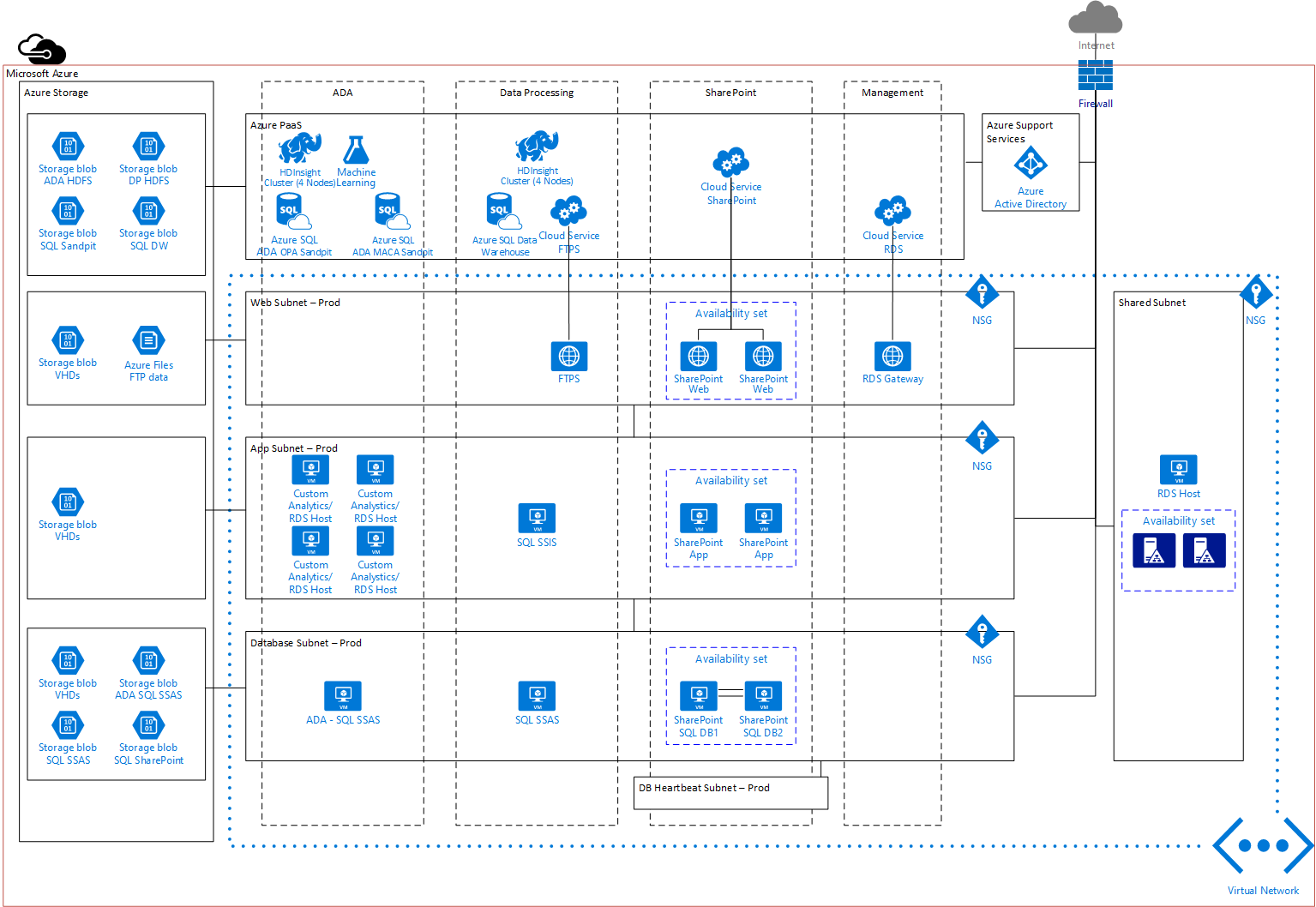
**42 DAP resource overview – test**

#### Production & Shared

The list of assigned resources for production is as follows.

| Azure Service | DAP Service | Resource | Description |
| --- | --- | --- | --- |
| Azure Support Services | Authentication | Azure Active Directory | PTV user authentication |
| Azure PaaS | ADA | HD Insight cluster – 4 nodes | HD Insight |
| Azure SQL | ADA sandpit |
| Machine Learning | Machine Learning |
| Data Processing | HD Insight cluster – 4 nodes | HD Insight |
| Azure SQL | Data warehouse database |
| Cloud service | FTPS data transfer NLB |
| SharePoint | Cloud service | SharePoint web NLB |
| Management | Cloud service | Remote desktop services gateway NLB |
| Azure VMs | ADA | A3 VM x 2 | Custom Analytics/RDS host |
| D3 VM x 2 | Custom Analytics/RDS host |
| A6 VM | SQL server SSAS |
| Data Processing | A2 VM | FTPS server |
| A3 VM | SQL server SSIS |
| D12 VM | SQL server SSAS |
| SharePoint | A3 VM x 2 | SharePoint web |
| A4 VM x 2 | SharePoint app |
| D12 VM x 2 | SharePoint database (SQL server cluster) |
| Shared | A2 VM | Remote desktop gateway (shared) |
| A4 VM | Remote desktop host (CGI support) |
| A2 VM x 2 | Active Directory |
| A4 VM | SCOM App |
| A4 VM | SCOM DB |
|  | Security | D2\_v2 VM | NAT |
|  |  | D3\_v2 VM | NextGen firewall |
| Azure storage – GRS | ADA | Storage blob | SQL sandpit |
| Storage blob | HD Insight HDFS |
| Storage blob | VM VHDs |
| Storage blob | SQL SSAS |
| Data Processing | Storage blob | SQL data warehouse |
| Storage blob | HD Insight HDFS |
| Storage blob | VM VHDs |
| Azure Files | FTP data |
| Storage blob | SQL SSIS |
| Storage blob | SQL SSAS |
| SharePoint | Storage blob | VM VHDs |
| Storage blob | SQL database SharePoint |
| Management | Storage blob | VM VHDs |
|  | Security | Storage blob | VM VHDs |

**43 DAP resources – production**



**44 DAP resource overview – production**

### Future Infrastructure Scalability Options

As there are many different components, the DAP can scale in multiple ways. The preference is to scale up first, and only scale out as a secondary measure. This information is provided as an indication of future scalability, subject to consideration as to the nature of the scalability issue being experienced.

* Storage performance can be improved by scaling out to more Storage Accounts. Each Storage Account can have 20,000 IOPS, with each VM restricted to 500 IOPS per disk.
* Storage Account capacity is limited to 500 TB per Storage Account, with 500 Storage Accounts available per subscription.
* SharePoint Web and Application servers can scale out by adding additional servers to the SharePoint farm. This will incur consumption costs of an additional server(s) running, and each new server will require additional SharePoint 2013 Enterprise Server licenses. Scaling up first is recommended. Microsoft do not support auto-scaling virtual machines for SharePoint, and any additional VMs must be built and added to the farm ahead of time (but can remain switched off until needed).
* Azure SQL Data Warehouse is scalable via the number of Data Warehouse Units (DWU) being consumed, and can be scaled at any time.
* Virtual Machines can be scaled up or down. An outage is required, which may take longer if the storage is being transferred from fast storage (solid state disks) to slower storage.
* Additional worker nodes may be added to HDInsight clusters. This may require the cluster to be torn down and rebuilt, but the cluster configuration can be saved prior.

## Operations Architecture

### Monitoring and Event Management

Monitoring and event management of the DAP solution will be provided at two levels:

* Azure VM monitoring
* Operations Insights

Faults that occur generate alerts that are visible through the Azure Portal, the OMS reports, the Security report (Power BI), and can be emailed or sent to an incident management tool (to be determined during move to support phase), allowing an incident to be raised to capture the fault.

Additionally, Data-level alerts from the ETL process are raised and presented through SharePoint (with the Alerting functionality of the DAP Portal) and/or email.

#### VM monitoring

Azure VM monitoring will be used to monitor the environment during the development phases of the project. Azure VM monitoring allows for both the collection of performance trends and the configuration of alert rules on a per VM basis.

For DAP, thresholds will be configured for key metrics with alerts sent to the CGI support teams via email.

Microsoft Operations Management Suite (OMS) will be implemented prior to DAP going into production.

#### Operational Insights

Operational Insights is part of Microsoft Operations Management Suite. This software-as-a-service solution provides for the collection, storage and analysis of log data from both on-premises and Azure systems.

The key capabilities include:

* Collect, combine, correlate, and visualise event and performance data
* Manage data centre capacity
* Ensure that the servers are up to date
* Track server configuration changes
* Review configurations for your workloads
* Know your malware situation

Currently Operations Insights is in preview. It is expected to be under general availability in time for the production DAP release.

### Simple Mail Transfer Protocol (SMTP Email)

The DAP needs to send email for operational monitoring and alerting purposes, as well as some notifications within the SharePoint environment. To provide this service, an Azure-certified cloud-based SMTP service, SendGrid is provided. As SharePoint cannot access the SendGrid mail API directly, an SMTP relay (version Windows Server 2012 R2) will be installed on the File Transfer server to provide this integration.

DAP is configured to use Azure SendGrid for SMTP based email for operational management and alerting (performance and service availability alerts), as well as for SharePoint notifications, such as workflow approval items. Emails are sent via the SMTP relay hosted on the File Transfer VM, and forwarded through to SendGrid for distribution. The SendGrid free tier licencing provides for up to 25,000 (twenty five thousand) emails per month. Additional volumes are available for a low fee, currently ~$10 per 20,000/month email blocks (this information was sourced from Azure/SendGrid).

Current email volumes in DAP are circa 30 (thirty) emails per month, with the system performing regular production data loads (myki and Smartrak data), which means DAP is currently operating in a production-like state from a data loading perspective.

If a problem was to occur where email alerts suddenly spiralled and exceeded the 25,000/month threshold, then alert emails would cease to be generated for the remainder of that month. The Azure SendGrid Dashboard provides a view on email volumes consumed at a point in time. SendGrid has an automated monitoring and alerting feature that provides alerts when predefined DAP administrator-defined email thresholds are exceeded. CGI has set up this monitoring and would raise an incident via the support process for investigation of the root case and to take corrective action.

The mitigation strategy if email was exceeded is to fix the root cause of the high level of alerts. To restore email capability for the remainder of that month, a licence for additional email volumes, with pricing in the order of $10 per 20,000 email blocks would be required. CGI also proposes periodic checks on SendGrid mail consumption as part of health checks, and if any inherent functionality is available within SendGrid to raise a volume-based threshold alert, then this would be configured.

There is no other impact to the business as a result of loss of SendGrid email, and users would be able to continue to access reports, data and the ADA area as normal.

### Performance Management and Capacity Management

As above, performance and capacity data will be collected through the Azure VM monitoring and Operational Insights. This will allow the VM performance and capacity to be reviewed periodically. Where necessary, recommendations for capacity adjustments will be made accordingly to maintain the performance of the solution.

This is an on-going process, and decisions to scale must be made in considerations of the additional cost that would apply.

Scaling can occur as a result of an alert due to sudden load, or as part of proactively viewing an increasing trend over time. The exact nature of the scaling activity must come from analysis of the resource that is under contention – for example, there is no point scaling the EDW to improve SharePoint response times.

Scaling options are defined in section 3.2, and typically involve scaling up virtual machines, scaling out SharePoint servers in the farm (incurring a licensing cost), or scaling up PaaS services.

### Data Protection

Data Protection will be provided by a number of mechanisms specific to each service type. This will include both short term and long term protection.

| Azure Service | Resource | Sort Term Protection | Long Term Protection |
| --- | --- | --- | --- |
| Azure PaaS | SQL DW | DB snapshots | - |
| SQL database | Automated Backups | Automated Backups |
| HD Insight | Azure Backups | Azure Backups |
| Azure IaaS | VM | VM snapshot | Azure Backup of VM |
| VM SQL Server | Dump to VM disk | Azure Backup of VM |

**45 Data protection approaches**

#### Azure PaaS

* **SQL Data Warehouse –** Microsoft provides snapshots of live databases every 8 hours that can be recovered for up to 7 days
* **SQL Database –** Microsoft provides business continuity features for SQL databases dependent on the service tier selected. These allow recovery to any restore point for up to 35 days
* **HD Insight –** HD insight will be protected using Azure Backup as discussed below

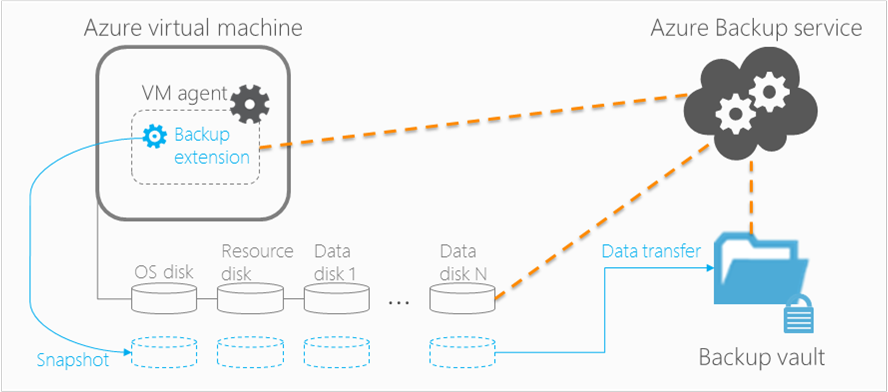
#### VM snapshot

VM snapshots allow rapid recovery of a VM to a previous state. In addition to automatic backups, VM snapshots will be manually used on an ad-hoc basis when changes are made to VMs (such as patching exercises), allowing rollback to a last known operating state.

#### Azure Backup

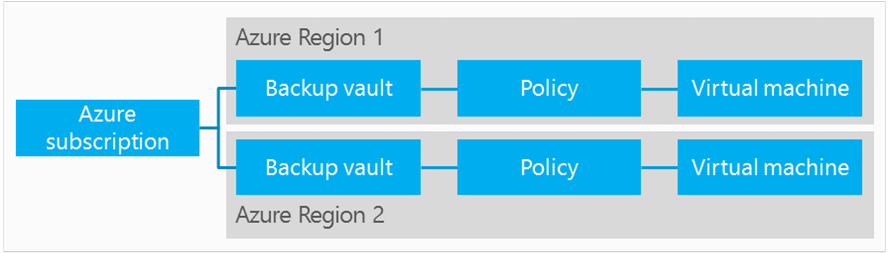
Azure backup provides policy based protection of Azure VMs. The Azure Backup service initiates the backup job at the scheduled time, and triggers the backup extension to take a snapshot. The backup extension coordinates with the in-guest VSS service to achieve consistency, and invokes the blob snapshot API of the Azure Storage service once consistency has been reached. This is done to get a consistent snapshot of the disks of the virtual machine, without having to shut it down.

After the snapshot has been taken, the data is transferred by the Azure Backup service to the backup vault. The service takes care of identifying and transferring only the blocks that have changed from the last backup – making the backups storage efficient. When the data transfer is completed, the snapshot is removed and a recovery point is created. This recovery point can be seen in the Azure management portal.



**46 Azure backup overview**

The vault is an entity that stores all the backups and recovery points that have been created over time. The vault also contains the backup policies that will be applied to the virtual machines being backed up.



**47 Azure backup vault**

For the DAP solution, Azure backup will be configured as follows:

* Vault region – Australia South-East
* Vault replication – geo redundant, maintains six copies of data across Australia South East and Australia East.

The Azure backup policy settings will be as follows.

| Item | Schedule | Retention |
| --- | --- | --- |
| Policy – Dev | Daily at 22:00 | 30 days |
| Weekly at Fri 22:00 | 5 weeks |
| Monthly | None |
| Policy – Test | Daily at 21:00 | 30 days |
| Weekly at Fri 21:00 | 5 weeks |
| Monthly last Fri 21:00 | None |
| Policy – Prod | Daily at 20:00 | 30 days |
| Weekly at Fri 20:00 | 5 weeks |
| Monthly last Fri 20:00 | 7 years |

**48 Azure backup policy settings**

### Disaster Recovery

All production and test services deployed with the DAP will be configured with resiliency, high availability and backup allowing for each individual service to be recovered from either configuration failure or user error. Disaster recovery will only be invoked when the Azure service itself cannot be provided from within the Australia South East region. In this scenario, and at Microsoft’s discretion, the Azure service will be failed over to the Australia East region.

In the event of Microsoft declaring a disaster in the Azure South-East region, the production system instances will be recovered in the Azure East region. This will be achieved via the following mechanisms.

| Resource Type | Resource | Recovery approach |
| --- | --- | --- |
| Azure PaaS | SQL database | Geo-replicated storage |
| HD Insight | Recreate cluster in DR region with geo-replicated configuration database |
| Azure IaaS | VM | Restart VM from geo-replicated storage |
| Azure Storage | Blob | Geo-replicated storage |
| Files | Geo-replicated storage |

**49 DAP resource recovery approach**

This approach has been chosen with the aim of recovering the DAP within one hour of the decision to fail over to the DR environment.

The Non-Functional Requirement for data loss is that no more than 20 minutes of data should be lost. This is supported by the DAP in that the sources appear to the Data Lake first, where six copies of the data is spread across two data centres, and the data can be reloaded by the Data Warehouse.

Any data stored in the Data Warehouse that is not stored in the Lake (such as ADA discoveries) are protected by transaction log backups, taken by Microsoft every 15 minutes.

As preparation for a DR failover, the virtual networks, subnets and load balancers will exist in the Australia East (Sydney) data centre.

In the event of a disaster that affects the DAP platform, and where Microsoft has not declared an entire Azure region emergency, each affected component will be restored via script from backup into the Australia East (Sydney) region. Due to the remote likelihood of a total Azure region failure, failing over to the Sydney data centre will be a scripted process, manually executed.

If Microsoft has declared an entire Azure region emergency, geo-replicated storage will be brought online by Microsoft in the Sydney data centre and VMs can be started directly from Sydney.

SaaS and DBaaS components are configured to automatically fail over to Sydney, and no action is required to bring these items back online.

Once the VMs have been brought back online, DNS changes will be necessary to re-route the web sites from the Australian SouthEast load balancers to the Australian East load balancers.

In future, Microsoft OMS Site Recovery will provide the ability to perform cloud-to-cloud DR, simplifying the process to bring up Azure resources in another Azure data centre.

### Disaster Recovery for Data Services

#### SSAS Tabular/Multidimensional Database

Recovering an Analysis Services database can be done in multiple ways, but the fastest is to restore from a backup of the Analysis Services database to a particular point in time, as this does not require re-processing of the data.

The backup and restore approach is also a valid technique for deploying a Tabular database to a server. e.g. deploying the tabular model to the ADA environment.

Backing up an Analysis Services database produces a backup file whose contents vary depending upon the storage mode used by the database objects. This difference in backup content results from the fact that each storage mode actually stores a different set of information.

The content of the backup file varies depending on the storage mode used in the Analysis Services database, described below:

|  |  |
| --- | --- |
| Storage Mode | Contents of Backup File |
| **Multidimensional MOLAP partitions and dimensions** | Metadata, source data, and aggregations |
| **Multidimensional HOLAP partitions and dimensions** | Metadata and aggregations |
| **Multidimensional ROLAP partitions and dimensions** | Metadata |
| **Tabular In-Memory Models** | Metadata and source data |
| **Tabular DirectQuery Models** | Metadata only |

#### Master Data Services Database

Master Data Services consists of a single database to store all master data and MDS system settings. Therefore, backing up the database can be integrated into a SQL Server backup and restore process.

MDSModelDeploy.exe can also be used to create a package containing the model objects and data.

#### Data Quality Services Database

Data Quality Services stores its information in three databases: DQS\_MAIN, DQS\_PROJECTS, and DQS\_STAGING\_DATA. Therefore, backing up the database can be integrated into a SQL Server backup and restore process.

DQSInstaller.exe can also be used to export all of the published knowledge bases from a Data Quality Server to a DQS backup file (.dqsb).

### Business Continuity

Once the DAP production systems have been recovered in the alternate Azure region, DAP users will be able to reconnect to the DAP services and resume normal operation. The automated update of DNS will make this a seamless process.

### Patching, Maintenance, Software Releases

The exact process for patching will be defined in operations documents, but the general process is defined below.

Patching will occur separately in each environment, Dev first, then Test, finally Prod, with sufficient time for testing patch impact before rolling out to the next environment. For each environment, the process is:

* Ensure all virtual machine backups have been completed successfully
* Apply patches via Windows Update to secondary/non-active machines, or in the case of an active/active configuration, half of the active nodes.
* Reboot and confirm successful installation
* Apply patches to remaining machines, reboot, and confirm successful installation.

In the event of a problem, servers can be rolled back via the initial snapshot backup.

This process allows each component to be patched with the rest of the system remaining available. There are a couple of components that do not have a redundant copy (such as jump hosts). These are designated non-critical servers, and may be offline for up to an hour during scheduled maintenance windows without affecting uptime of the rest of the system.

Deploying new features to the DAP will follow a similar process of testing in Dev and Test prior to general release.

# Security Architecture

At a high level, the security architecture is as follows, with all user access flowing through HTTPS (only internal systems, such as ADA, will have access to TDS).

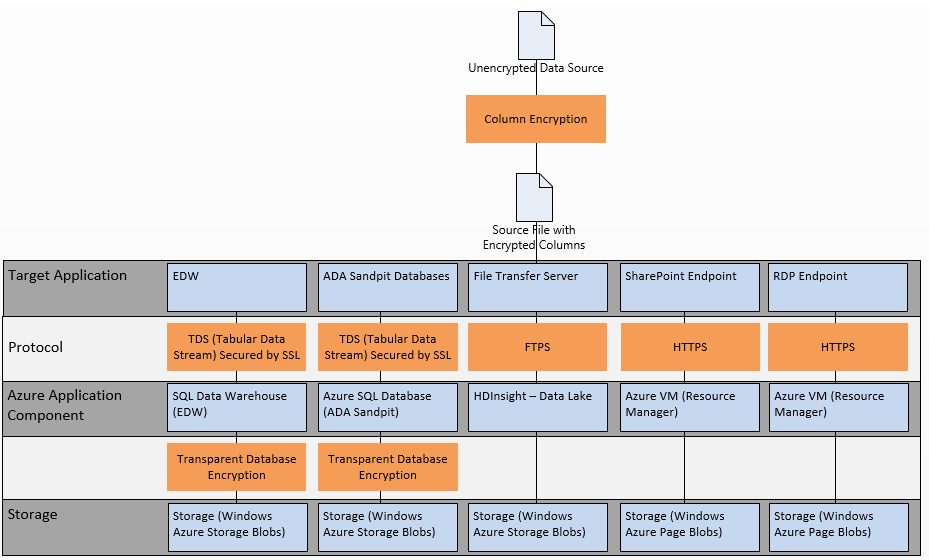


Figure – Security Architecture

For a detailed description of how security requirements are handled against the ASD 35, OWASP, SDLC and ISM security checklists, see the DAP Security Assessment Report.

## Network Security

There is a single virtual network with multiple virtual subnets. From an Azure technical perspective, there is no security difference between multiple virtual networks joined together compared with multiple virtual subnets protected by Network Security Groups (NSGs). For more information, see the Infrastructure technical design.

Public Network Endpoints are provided by Azure Network Load Balancers (NLB). The NLBs are:

|  |  |  |  |
| --- | --- | --- | --- |
| NLB | Production | Test | Development |
| Data Transfer Endpoint (FTPS) | Yes | Yes | Yes |
| SharePoint Endpoint | Yes | Yes | Yes |
| Remote Desktop Endpoint (Shared Infrastructure) | Yes | - | - |

Each endpoint is monitored by the Azure Network Operations Centre, with automatic detection of denial of service attacks (from either inside or outside of Azure), and log analysis of activity.

### Web Application Firewall (WAF)

Web application firewalls (WAF) control access to applications or services using deep protocol inspection to detect anomalies for web based services. The WAF is designed to detect, drop or scrub malicious or malformed content as per the OWASP Top Ten List ([www.owasp.org](http://www.owasp.org)).

To provide an additional layer of security to the DAP environment an F5 WAF solution has been deployed in front of the Test, Production and RDP gateways.

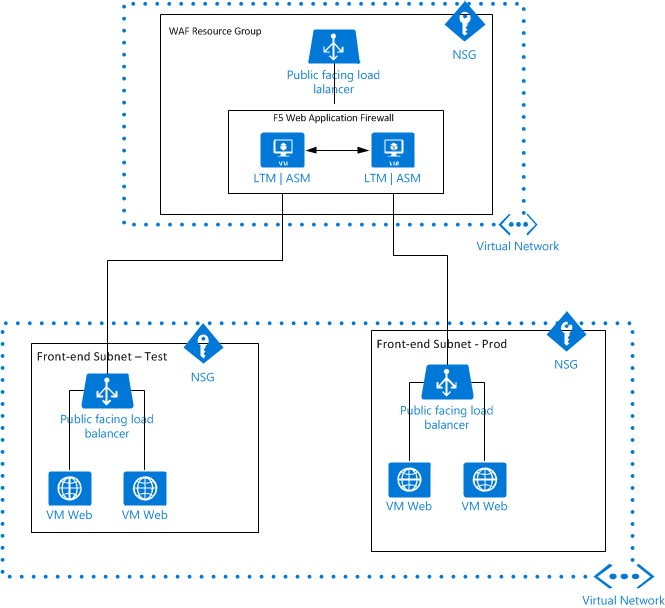


Figure – WAF Deployment Overview

F5 WAF is a Microsoft Azure Partner solution, offering virtual turnkey deployment into an environment providing immediate protection to web applications deployed behind it including OWASP Top Ten.

F5 WAF is deployed via the Azure Security Centre where additional applications can be deployed or removed from firewalling simply and efficiently.

Monitoring is also integrated into the Azure Security Centre.

The deployment consists of two VMs deployed as an Active-Active HA pair each with a 200mbps throughput license. In this configuration, the devices load balance and provide a total throughput of 400mbps whilst both VMs are operational. In the event of a failure, all traffic will be load balanced across to the available VM but throughput available will be halved to 200mbps.

### Access Control Intermediary (Proxy) for ADA environment

The ADA environment allows users to install software on the VM workstations and have database access to support analytical use cases. In order to prevent potential misuse of this platform and to provide a degree of protection for the environment a NextGeneration firewall has been deployed into the DAP environment.

The ADA environment has been segmented away from the production application environment so that firewall policies only apply to ADA currently. The solution chosen does not currently support high availability within Azure, so it was vital that the impact of an outage on the firewall was limited.

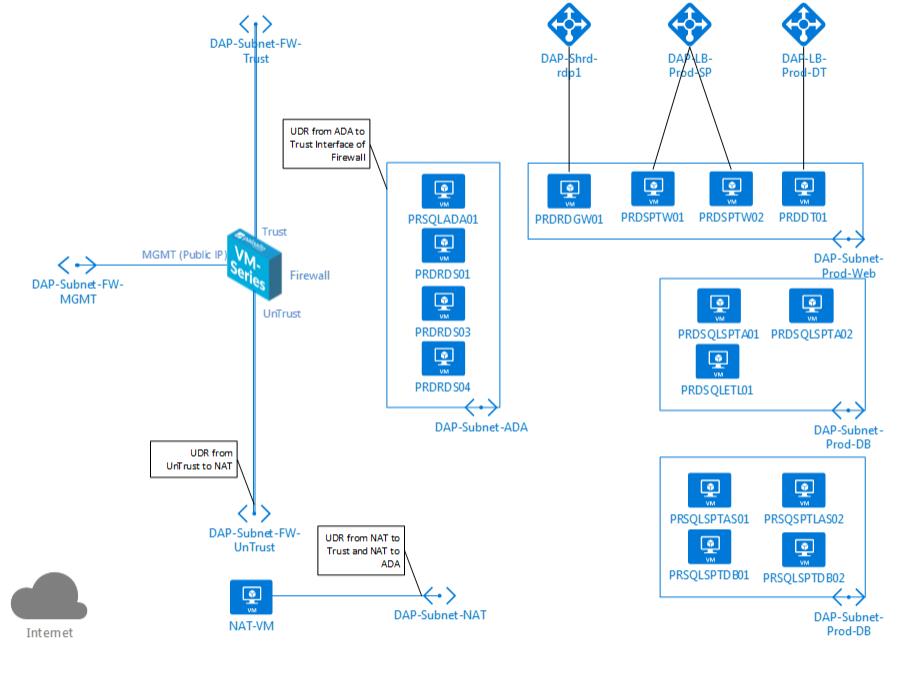
By default, all devices in Azure have internet access as well as access to all other subnets within the environment. User defined routes (UDR) control traffic flow within Azure. UDRs have been defined to control the flow of traffic out of and into the ADA subnet for inbound and outbound internet traffic to force all internet traffic to the firewall for access control and content inspection purposes.

URL content filtering allows or restricts access to certain sites based on pre-defined categories. This allows businesses to define what sites users are permitted to have access to, with the aim of allowing legitimate internet access for business purposes whilst denying access to illegal, inappropriate or malicious content.

URL content filtering has been applied to the ADA environment along with access policies that control internet access at the application layer. Policies are based on positive enforcement principles in that specific content is permitted with the rest being denied by default.

Logging of content allows visibility of permitted and denied traffic, along with threat analysis.

Policy will develop over the lifetime of the environment and consideration may be given to further firewalling of ADA and the internal environment.

Figure 52 – ADA Firewall Overview

## System Security

### Azure Subscription Access

Administrators and Co-administrators of the Azure subscription can do anything in the Azure environment, including provision new services, decommission services, and access all data.

Few people will require access at this level, as most work will be performed on services that have been built.

### Virtual Machines Management

Access to virtual machines in the DAP is based on Active Directory permissions and granted via group membership.

The following groups will be created. All groups will be split into Development, Test, and Production categories.

|  |  |
| --- | --- |
| Group | Access |
| Domain Administrators | All access to the DAP Active Directory (on-premises version) in Azure. Administrator-level access to all servers. |
| SQL Server DBAs | Administrator-level access to all SQL Server servers. |
| SharePoint Administrators | Administrator-level access to all SharePoint servers (excluding database servers). |
| Developers | Users on all development servers |
| ADA Administrators | Administrator-level access to all ADA servers. |
| ADA Users | Users on all development servers |

### Logging

Logs created by virtual machines are only accessible to Domain Administrators and local Administrators for that server.

Audit logs created by PaaS SQL Server (including Azure SQL DB and Azure SQL Data Warehouse) are targeted towards a storage account, and only Azure Administrators will have access to this storage account or the ability to start, stop, or modify logging functionality.

SharePoint provides logging in the form of Web Analytics logs, and W3SVC logs, and access to these logs is available only by administrators of the SharePoint servers.

As part of the security enhancements as a result of the DAP Security Assessment, a log monitoring solution may be provisioned to centrally manage and protect logs as part of a Security Information and Event Management (SIEM) tool. The SIEM tool should assist with protecting from log tampering by use of remote logging or log replication. Additionally an Intrusion Detection System (IDS) may be implemented to assist in monitoring changes to the environment.

## Data Security

The Microsoft Azure platform is IRAP (Australian Signals Directorate - InfoSec Registered Assessors Program) certified, and the development of the DAP will be undertaken alongside analysis of:

* the Privacy and Data Protection Act 2014
* Victorian Government Information Management Program
* Victorian Government Privacy Standards

No passwords or storage keys are to be stored in application code, and must be encrypted.

### Data Lake - HDInsight

Access to the data stored in the Data Lake is limited to those that have a Shared Access Signature (SAS) key to read or read/write into the Storage Account or Storage Container.

At the time of writing, HDInsight or Azure Storage Blobs does not support transparent encryption, but Microsoft advise this feature will be available in the next 12 months with the General Availability of Azure Data Lake (ADL), which is the preferred target state for the DAP Data Lake due to the encryption and Active Directory integration for file-level security. Until ADL is available, data stored in the DAP Data Lake will be unencrypted files, with sensitive columns encrypted within the file.

These columns should be encrypted as close to the source data as possible (on initial extract), and unencrypted on use.

Encrypted columns include:

|  |  |  |
| --- | --- | --- |
| Data Source | Column | Owner |
| myki | myki PAN number (Card number) | Chris Maloney |
| myki | Driver ID | Chris Maloney |
| myki | Driver Shift ID | Chris Maloney |
| myki | Terminal Group Location | Chris Maloney |
| myki | Service Location | Chris Maloney |
| myki | Customer Title | Chris Maloney |
| myki | Suburb | Chris Maloney |
| myki | Postcode | Chris Maloney |
| Smartrak | Company Name | Richard Bennell |
| Smartrak | Vehicle Registration | Richard Bennell |
| Smartrak | Depot Name | Richard Bennell |
| Smartrak | Driver ID | Richard Bennell |
| Smartrak | Remote ID | Richard Bennell |
| Smartrak | Run ID (Shift) | Richard Bennell |

Myki data is encrypted on the source server, as part of the extract, while Smartrak data is encrypted after downloading the JSON file (as the Smartrak Webservice does not encrypt data above the HTTPS protocol), prior to conversion to CSV and storing to the Data Lake.

### Azure SQL Database and Azure SQL Data Warehouse

All Azure SQL Database and Azure SQL Data Warehouse will have the Transparent Data Encryption feature installed. This encrypts the data on disk, and each page is decrypted as the page is loaded into memory. This ensures that only database servers that have a copy of the encryption key can view the data.

Azure SQL server Auditing will be enabled to log all failed events, and logins. This is written to the Audit storage account, and is only accessible by Azure subscription Administrators/Co-Administrators.

Azure SQL Database Threat Detection is a new feature that detects anomalous database activities which may indicate a common threat like SQL injection attacks. The Public Preview of this feature started in November 2015, and will be enabled when Generally Available in Australia.

When reading data from the Data Lake through the PolyBase layer, the Azure SQL Data Warehouse is protected by segregation. PolyBase executes on a Java Virtual Machine segregated from the SQL DW, and only if the JVM successfully returns a dataset will it load any data into the Data Warehouse.

#### Access to Data stored in Azure SQL Data Warehouse

The Business Vault and data marts are mainly accessed through SSIS, SSAS and SSRS.

Only DAP (database) administrators will have direct and full access to the Raw Data Vault. Direct access to the database for any other user is not recommended and must be treated with caution. However if there is a justifiable need to allow direct access to these databases then it must conform to these rules:

* Permissions that are granted to a role must be reviewed periodically
* Membership of the roles must be verified and extended periodically.

Note: Allowing users to have write access to the DAP databases may compromise data integrity.

All user access to SQL DW and SQL DB is performed via SQL logins, and each user will be granted a login, and be assigned a role in the database. No user will be granted access directly to specific database resources (tables, views, procedures, etc).

### Power BI Data

As mentioned in section 10.1.2 - Power BI, Power BI is hosted outside of Australia. Report and Visualisation definitions are stored on the Power BI servers, as is any uploaded data source. The DAP data sources, available to any user with a Power BI Pro subscription are stored in Australia, and the data is only stored in Australia.

Power BI is an organisation-based solution, and reports cannot currently be shared with people who are members of another organisation.

### Azure Machine Learning

As Azure ML is currently only available outside of Australia (but utilising Storage Accounts hosted in Australia), it may not be appropriate for all of PTV’s data. Data is thus stored in Australia, but processed in-flight (not stored) by the Azure ML services.

Data sets to be used with Azure ML should be examined closely to ensure they do not contain sensitive information, and any data set that cannot be processed offshore by Azure ML due to its sensitivity should be processed using local tools in the ADA environment.

### SQL Server Analysis Services – Semantic Layer

There is a business requirement to restrict access at the data row level – e.g. Operators can only view information relating to their area. This will be achieved by implementing dynamic row level security within the Tabular Model.

Implementing row level security within Tabular In-Memory models requires:

* A security access table within the model containing the Windows user names of individuals mapped to the allowed data (e.g. Operators).
* A user role, created in SQL Server Data Tools using the Role Manager. This role will include a row filter (DAX expression) defining which rows are visible to users.

There are challenges in implementing access control at this level some of them are:

* Ensuring the security access table contains accurate and current data at all time. This will require implementing a process that gets the users, roles and roles membership.
* Managing the impact on performance.

**Note:** Row level security will need to be provided by thebackend database in Tabular models with DirectQuery enabled. DirectQuery allows the Tabular model to leave detailed data in the database and utilise Azure SQL Data Warehouse to perform the heavy lifting of detailed data, resulting in a leaner model that requires less memory.

### Storage Accounts and Containers

The DAP will use many different Azure storage accounts and containers. By default, all storage accounts are private, and can only be accessed from within the Azure subscription. Storage accounts and containers can be opened up to allow read-only and read-write access to users that have been allocated a Shared Access Signature (SAS) – a key that allows delegation of access to specific storage account resources. SAS keys can be revoked at any time. Storage accounts can also be designated as public, where no SAS key is required, and anyone can access files stored there. This is not recommended for the DAP solution.

The following storage account types will be changed from the default:

|  |  |
| --- | --- |
| Purpose | Description |
| Incoming Files | Files uploaded directly to the Data Lake via the Azure Blob storage API will require a read/write SAS key. Each external partner will have their own container, and will only have access to that container. |
| Data Lake access to ADA users | As described in section 7.1.1 - Data Lake Storage Accounts and Containers, there are storage accounts and containers for each data source. This allows, for example, only MACA users to have access to a SAS key for the myki ticketing data. |

All data stored in Azure is at risk during hardware disposal. Hardware disposal is governed by Microsoft’s Global Foundation Services policy that details procedures for wiping or destroying drives and securely disposing of hardware from Windows Azure data centers. These policies are designed to minimize the risks of anyone obtaining hardware containing data. More information is available from <http://blogs.technet.com/b/gfs/archive/2009/05/27/securing-microsoft-s-cloud-infrastructure.aspx>

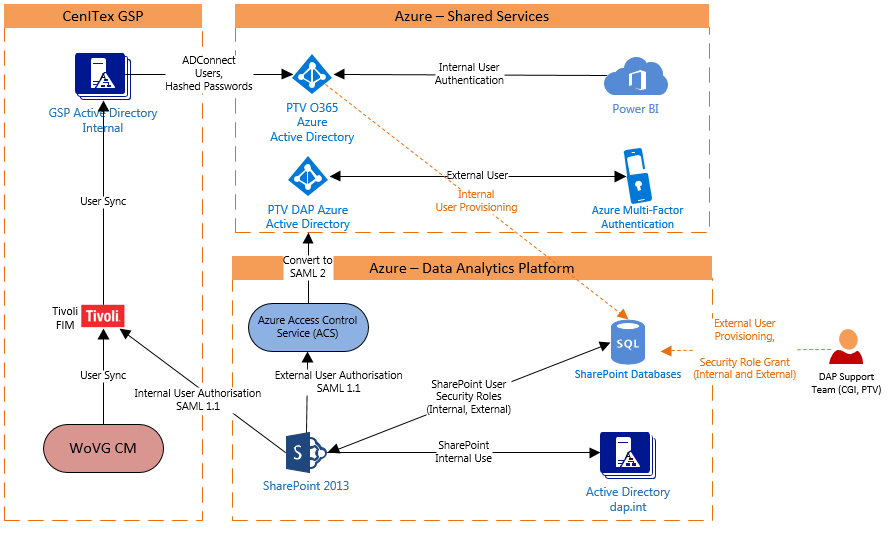
## Application Security

All application service accounts follow the principle of the fewest required privileges to complete the task.

### SharePoint Authentication

There are two main authentication targets provided in the DAP:

* PTV staff are authenticated against Tivoli FIM
* External users are authenticated against the PTV DAP Azure Active Directory



**53 SharePoint Authentication Mechanisms**

SharePoint permissions are controlled via roles. Only users that have been provisioned in SharePoint will be associated with a role and be able to access content. The DAP support team will grant access to users, and create external user accounts. Other government agencies that have access to the DAP will also authenticate against Tivoli FIM, after an account has been created inside SharePoint. Exact group membership definitions have not yet been defined, but will be defined as part of the SharePoint User Interface design document.

SharePoint will present multiple authentication types, for Internal and External users.

PTV users coming from a trusted location/browser/IP (on the Government Shared Platform, GSP) will experience single sign on through CenITex’s Tivoli FIM. PTV users coming from an untrusted location/browser/IP are directed to the CenITex Tivoli login page, which requests a username, password and multi-factor authentication. Any VicGov user that does not have a CenITex multi-factor authentication token will be unable to access the system remotely, and must instead request external access from CenITex, or log in via a CenITex GSP host.

External users are authenticated against the PTV DAP Azure AD, with multi-factor authentication. Multi-factor authentication is provided by Azure AD, and consists of an SMS sent to the user’s mobile phone.

For information on SharePoint application security, see Appendix A (SharePoint).

### SQL Server Analysis Services (SSAS)

SSAS uses a role-based authorisation model to control access to cubes, dimension data and other SSAS objects. It allows assignment of AD roles and groups to individual SSAS objects. The DAP solution will use SSAS’s in-built authorisation to control access to SSAS objects and models.

Connection of the SSAS objects to the data marts will be through a specific service account with limited read-only access.

### SQL Server Integration Services (SSIS)

ETL components that are developed in SSIS will need to access the DAP Staging database and Data Vault database as well as the data marts. These components will access these databases with a specific service account for read and write purposes. The service account used to connect to the EDW is only for running ETLs, and not for general use. This account is one of the very few to have write access to the Data Vault, Business Vault and Data Mart schemas.

### Master Data Services (MDS)

Selected users (Data Stewards) will have access to Master Data Services to maintain the reference data (tables) for the DAP solution. Master Data Services uses role based authorisation based on local or Active Directory domain users and groups. This allows a granular level of detail when determining the entities and data a user can access.

Note: Security settings applied in the Master Data Services web application are also applied to the Add-in for Excel.

### Certificates

Certificates are required in the DAP for the following purposes:

* HTTPS communication for SharePoint
* FTPS for file transfer
* Access to the Remote Desktop Services Gateway over HTTPS

As per the PTV Encryption Standard, applications utilising HTTPS or FTPS must use a certificate, issued by a trusted certificate authority, which clearly identifies the authenticity of the application.

### Application Whitelisting

Application whitelisting is a security technique whereby only a limited number of programs are allowed to run whilst all others are blocked by default.

Due to non-privileged access to the ADA platform, the ADA remote desktop VMs were assessed as being the most vulnerable to malicious programs being inadvertently executed.

The Microsoft operating system has security controls built-in that can be configured to block potentially malicious content from being executed.

#### Microsoft AppLocker

AppLocker is an application control feature that allows the creation of rules either allowing or denying applications from being run based on unique identities of files and by specifying which users can run these applications.

AppLocker has been deployed to the ADA RDP VMs via group policy and has been configured to only allow the following;

* All users can execute files in the following;
  + All files located in the Program Files folder
  + All files located in the Windows Folder except Windows\Temp
* All users can install;
  + All digitally signed Windows Installer Files
  + All windows Installer files in %systemdrive%\Windows\Installer
* All users can execute scripts;
  + All located in the Program Files directory
  + All scripts located in the Windows folder except Windows\Temp
  + All trusted scripts in the [\\dap\SYSVOL\dap.int\\*](file:///\\dap\SYSVOL\dap.int\*)
  + All trusted scripts in the [\\dap\NETLOGON\\*](file:///\\dap\NETLOGON\*)

Deployment of AppLocker deployment is performed by moving the appropriate AD computer objects into the dap.int\AppLocker Restricted AD OU.

Additional security restrictions may restrict these operations further.

## User Client Security

All user access is authenticated against an Active Directory implementation, whether by an on-premises AD in the cloud, or an Azure Active Directory.

The only exceptions are for Power Users that have access to:

* Storage Accounts and Containers (secured by use of a Shared Access Signature)
* Direct access to Azure SQL Database and Azure SQL Data Warehouse, which require a SQL login, stored in the database
* Access to the HDInsight clusters

### DAP AD Account Management

The DAP Active Directory supplies accounts for ADA users. Using the Active Directory console, these accounts can be enabled and disabled at any time by the CGI DAP Support team, and this takes effect immediately.

The following password complexity rules apply for DAP AD user accounts:

* Not contain the user's account name or parts of the user's full name that exceed two consecutive characters
  + Be at least twelve(12) characters in length
  + Contain characters from three of the following four categories:
    - English uppercase characters (A through Z)
    - English lowercase characters (a through z)
    - Base 10 digits (0 through 9)
    - Non-alphabetic characters (for example, !, $, #, %)

Complexity requirements are enforced when passwords are changed or created. Passwords must age for at least 24 hours before they can be changed again by the user.

Only the CGI DAP Support team has access to modify Active Directory. The only action a user can take is to change their password.

Note that the passwords used by FTPS users cannot be changed by the users. As an operational measure, these passwords should be changed and communicated to NTP and Hoban at least quarterly.

### Source Code Access

All source code is stored in PTV’s Bitbucket repository, which some ADA users may have access to. Additionally, stored procedures within the EDW and SSIS packages stored on the ETL server can be considered an accessible source. Some ADA users may be granted access to view the code of stored procedures, or access Bitbucket, or access the SSIS server, but this is by exception only, and the code is not publically accessible.

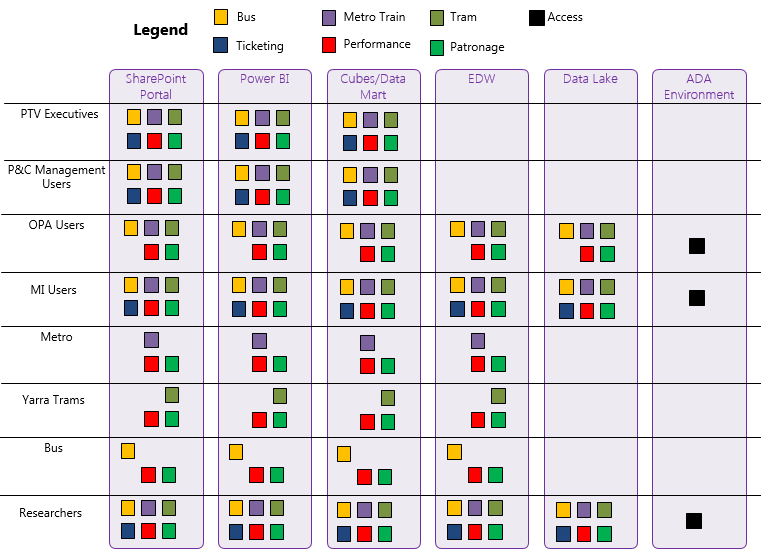
## User Classes

The following user classes have been identified in the DAP. All users have access to SharePoint, but may be restricted from viewing some reports based on the reports’ classification.

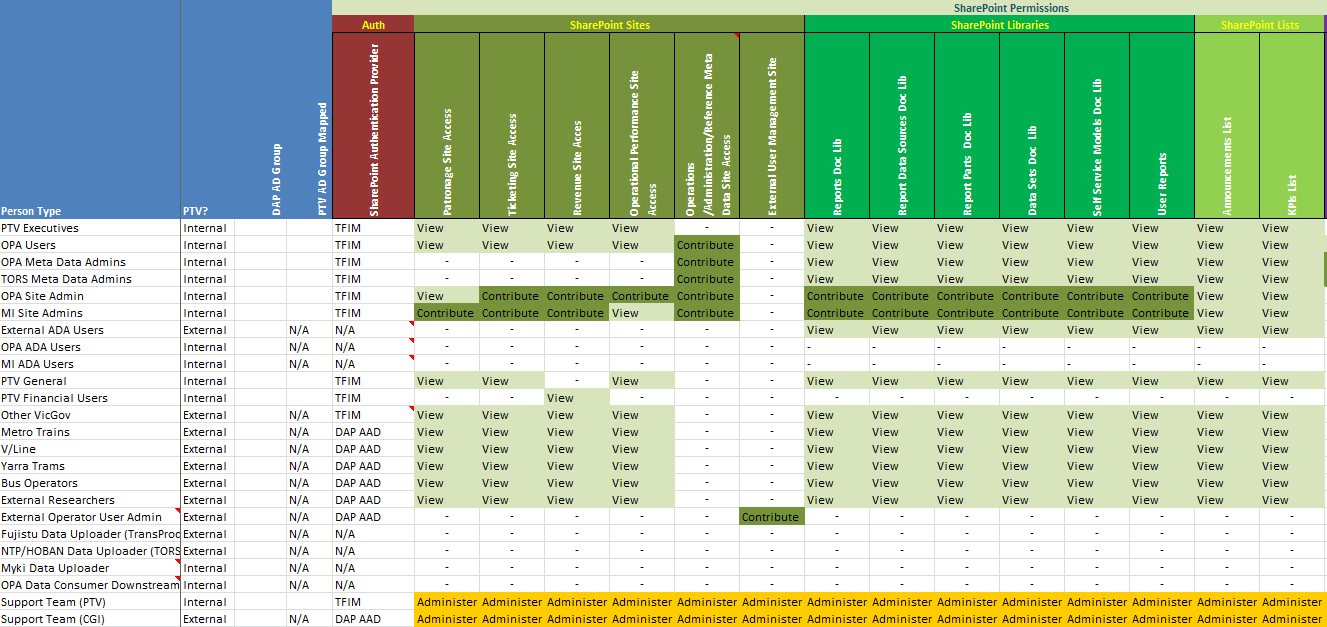
|  |  |  |  |
| --- | --- | --- | --- |
| User Class | Description | Est. Number of Users | Est. Concurrent Users |
| Azure Administrators | Responsible for managing the Azure infrastructure, including billing and consumption, and access to the Azure Management Portal to provision new Azure services and maintain existing ones | 5 | 2 |
| Developers | Developers, mostly within CGI, and some PTV users that maintain and develop the EDW and its components. | 25 | 15 |
| Reporting Users (PTV) | Reporting users within PTV have access to SSRS reports presented through SharePoint | 400 | 40 |
| Reporting Users (External) | External Reporting Users from PTV’s partners, such as Metro Trains, Yarra Trams, Bus companies have access to SSRS reports through SharePoint. | 50 | 10 |
| Reporting Power Users | Reporting Power Users have access to Power BI Pro, and can create and view Power BI reports. | Year 1: 175 Year 3: 235 | 20 |
| ADA Users (PTV) | ADA users within PTV have full access to the ADA area to analyse data, and access to data stored in the Data Lake that matches their jurisdiction. | 20 | 10 |
| ADA Users (External) | Some external users can be provisioned access to the ADA area, and granted limited read-only access to some data sets hosted in the Data Lake. | 5 | 2 |

## User Permissions Matrix

At a high level, the following permissions apply to Reporting users and the data they can see. This affects what is visible within the reporting tools, and the level of data that can be viewed during self-service BI:



The following except shows the user classes and their permissions within the DAP. For the full document, please see the attached DAP Permissions Matrix in section 13.3 Attachments. This is considered a living document, and permissions and classes may change as the solution is built.



## User Maintenance Processes

The DAP allows for two main classes of user to access the system – Internal PTV users and External users. This section details how user onboarding and offboarding should be managed.

### Internal PTV Users

Generally, Internal PTV users do not need to be granted access to the DAP. A SharePoint user profile is automatically created through SharePoint once Tivoli has authenticated their credentials for the first time. The SAML token provided by Tivoli includes an authorisation for PTV staff (based on their email address being in the ptv.vic.gov.au domain) to be in a group named “SharePointReader”, and the DAP portal maps this group for internal PTV basic access. Additional access to areas within the SharePoint portal will be provisioned by the DAP Support team. A user can request this access via the contact form provided in SharePoint, or a direct email to the PTV DAP Support team, who will consider the request and provision the access.

Other VicGov users that do not have a PTV email address are not granted automatic access, and must be provisioned manually, granting access to the appropriate roles for their required access. Non-PTV VicGov users will request access using a similar process to External Users (see next section), but with no requirement to provision an Azure AD account.

By the same token, when a VicGov user leaves the organisation, their account will be automatically deactivated in Tivoli, and SharePoint will be unable to authenticate the user, removing their access to the DAP.

If a new Internal PTV user requires Power BI Pro access, they will need to be assigned a Power BI Pro license through PTV’s view of the Microsoft Office365 portal.

Requesting access to the DAP for internal and other VicGov users is performed via the Government Identity Provisioning Service (GIPS) system which handles approval and routes actions to the DAP support team to provision the requested level of access. A page hosted in DAP’s SharePoint is linked to from GIPS to supply information to the requestor on the subject of DAP access levels.

Users that have access to ADA will need to have their accounts deactivated manually by the DAP support team. This request is automatically forwarded to the DAP support team via GIPS upon user deactivation.

### External Users

To gain access to the DAP, External users must be added manually into the PTV DAP Azure Active Directory and granted group membership to access specific reports. While manual, this process can be streamlined via scripting. While the process may undergo some changes over the first 12 months of operation to optimise the process, the process will be:

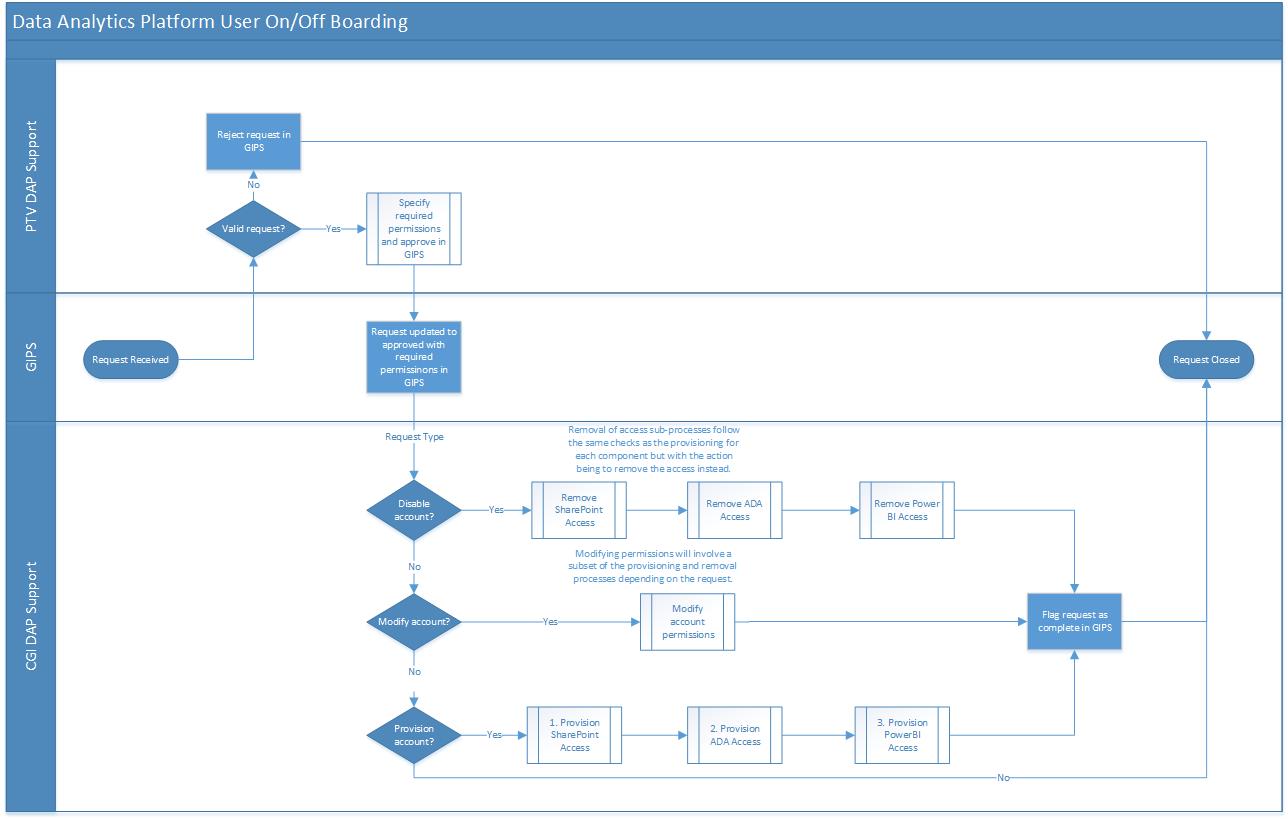
1. An external user requests access to the DAP via a form in SharePoint. This ensures that the user requesting access has been given the right to request access for new users. The required fields for a new user are: First Name, Last Name, Email Address, and a mobile phone number to be used for Multi-Factor Authentication.
2. A member of PTV’s DAP Support Team validates the user’s identity with a known approver within the external organisation. The exact criteria for this process will be refined as part of transition to go-live, with a minimum of ensuring the email address of the new user matches the domain of the requesting organisation (e.g., Metro Trains users must have a valid @metrotrains.com.au email address).
3. The user account is created, including setting a mobile phone number for multi-factor authentication, and initial credential details are forwarded to the user. The initial password must be changed upon first login, and as the phone number for multi-factor authentication has been configured by the DAP Support team, only a person in possession of that mobile phone will be able to use the credentials.
4. Upon first login, the user enters their temporary password, and must immediately set a new password before Azure AD will authorise the SAML token for SharePoint.

Offboarding an external user is difficult, as PTV may not be made aware of an external user leaving their organisation. The following process minimises the exposure to inactive accounts:

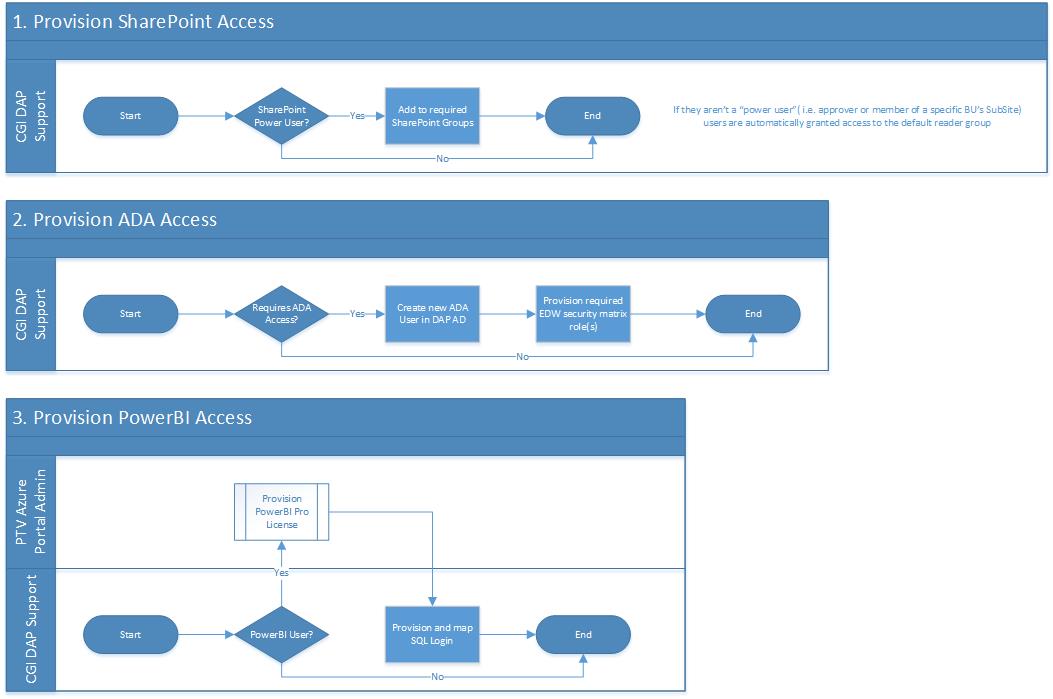
1. SharePoint Web Analytics is enabled, logging the activity of all users on the SharePoint portal.
2. Weekly, the SharePoint Web Analytics are reviewed via an automated process, and for each user name, the most recent access date/time is extracted and stored.
3. On a weekly basis, the external users are compared to their most recent login time, and any that have not logged into the portal within the past 4 weeks is emailed a reminder to log in (which will reset the counter once SharePoint Web Analytics captures the new activity). If no action is taken, the user will be reminded twice more. This is an automated process.
4. Any user that has not accessed the system within 6 weeks will have their account disabled, and will be unable to access the portal. A user in this state will contact the DAP Support team via email to request the account be unlocked.
5. Any accounts that have been disabled for 16 weeks are removed.

After 10 unsuccessful logon attempts (wrong password), the user will need to solve a CAPTCHA dialog as part of logon. After a further 10 unsuccessful logon attempts (wrong password) and correct solving of the CAPTCHA dialog, the user will be locked out for a time period. Further incorrect passwords will result in an exponential increase in the lockout time period.A future phase of the DAP may implement self-service registration with the ability to set a list of valid partner organisations (based on email address domain), and the ability to set specific default access levels for each partner organisation, and the ability to request additional levels of access. The workflow for approving these requests can be handled automatically, or included a manual approval step.

### User Onboarding/Offboarding Overall Process Flow



### User Onboarding/Offboarding Sub-Process Flow



# Glossary, Appendix & Attachments

## Glossary

| Term /Acronym | Definition / Expansion / Description |
| --- | --- |
| DAP | Data Analytics Platform |
| FTPS | File Transfer Protocol Secured by SSL. Not SFTP, which is based on Secure Shell (SSH). |
| Data Vault | Data Vault is a data warehouse modelling technique that allows for a flexible schema, and promotes a data model that is able to agilely adapt to new data sources or changes without affecting other parts of the EDW. |
| EDW | Enterprise Data Warehouse |
| R | The R programming language, used for statistical analysis |
| PolyBase | Inter-datasource access layer provided by Azure SQL Data Warehouse to query homogenous and heterogeneous data sources directly through the T-SQL language. |
| WASB | Windows Azure Storage Blob |
| T-SQL | Transact-SQL, Microsoft’s implementation of the SQL standard |

**54 Glossary**

## Appendix

| Appendix | Title |
| --- | --- |
|  |  |

55 Appendix

## Attachments

| Attachment | Title |
| --- | --- |
|  | NFR Mapping Spreadsheet |
|  | DAP Infrastructure Configuration |
|  | DAP Permission Matrix |
|  | DAP SharePoint Architecture Details |
|  | DAP Dimensional Bus Matrix |
|  | DAP Estimated Consumption |
|  | Risk Assessment for SA-R5 |

56 Attachment